

The Efficacy of Propensity Score Matching in Bias Reduction with Limited Sample Sizes

Stephani Howarter, M.S.

University of Kansas

Submitted to the graduate degree program in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Vicki Peyton, Ph.D.

Co-Chairperson William Skorupski, Ed.D.

Bruce Frey, Ph.D.

Neal Kingston, Ph.D.

Barbara Phipps, Ph.D.

Date Defended: December 7, 2015

The Dissertation Committee for Stephani Howarter
certifies that this is the approved version of the following dissertation:

The Efficacy of Propensity Score Matching in Bias Reduction with Limited Sample Sizes

Chairperson Vicki Peyton, Ph.D.

Co-Chairperson William Skorupski, Ed.D.

Date approved: December 7, 2015

Abstract

The current literature on propensity score matching is missing imperative information for educational researchers regarding the practical implications of utilizing this method with limited sample sizes. The purpose of this study was to evaluate the effectiveness of propensity score matching when limited by sample sizes of 500,400, 300, and 200 as determined by a reduction in bias using both real and simulated data. Further effort was made to determine the optimal selection of covariates and caliper width with these limited sample sizes. Participants were selected without replacement and matched one-to-one using the nearest neighbor technique in the MatchIt package in the R software program. Contrary to the hypothesis that with reduction in sample size the balance improvement would drop below what is considered effective bias reduction, the reduction in bias was greater than 96.77% for all conditions of sample size and caliper width. A Monte Carlo simulation was created based on the real dataset to assess covariate selection with the same limitations in sample size and a set caliper width of 0.6. For all replications, the mean balance improvement was best for the covariate relationship magnitude strong_none (strong relationship to DV_no relationship to treatment) and worst for the relationship mod_strong (moderate relationship to DV_strong relationship to treatment). Only the covariate relationship strong_none was able to be deemed effective matching for all sample sizes. Findings suggest that propensity score matching can be effective at reducing bias with sample sizes as small as 200 and caliper widths as wide as 0.6. Ideal covariates are those that are strongly related to the outcome variable and only weakly or moderately related to treatment when sample sizes are limited.

Keywords: Propensity Score Matching, Sample Size, MatchIt

Acknowledgements

I would like to sincerely thank the support of my entire committee members, Dr. Vicki Peyton, Dr. William Skorupski, Dr. Bruce Frey, Dr. Neal Kingston, and Dr. Barbara Phipps not only for this dissertation, but throughout my educational career. Their knowledge and passion for the field of measurement has served as an inspiration to me. I would especially like to thank my advisor and chair Dr. Vicki Peyton and my co-chair Dr. William Skorupski; without their belief in me and encouraging words this never would have happened.

I would also like to thank my colleagues that challenged me and provided encouragement, Susan Gillmor and Jessica Loughran. Thanks for the laughs and for being good friends. I also appreciate the time and assistance with coding from Jared Harpole. Thank you for being an R master.

Finally, I could not have completed this journey without the love and support of my family. To my parents, who have always believed in me and supported my decisions; my drive and love for education is a direct influence from you. To my partner in crime, Kelly, thanks for always being there for me. You are a blessing and you keep me sane. To my husband, Mark, thank you for everything—from taking over in my absence to listening to hours of talk about code and simulations. Your support and tireless optimism helped in so many ways. For Emi and Lydia, I really did this for you. All my love.

Table of Contents

Chapter One - Introduction	1
Chapter Two - Review of Literature	4
Propensity Score Analysis	6
Assumptions.....	7
The Ignorable Treatment Assignment Assumption	7
The Stable Unit Treatment Value Assumption	9
Steps for Propensity Score Matching.....	12
Covariate Selection	13
Calculate the Propensity Score	15
Matching Analysis	17
Determining the distance for matched pairs.....	18
Mahalanobis Distance Measure	18
Matching within calipers.....	19
Choosing an algorithm for matching	20
Nearest available matching.....	20
Optimal Matching	21
The structure of the matched set	21
Postmatch Analysis	24
Assessing the Matching Quality	24
t-test of significance.....	26
Standardized Bias.....	28
Plots and Histograms	29
Propensity Score Summary Statistics.....	29
Estimating the treatment effect	30
Advantages.....	31
Limitations	32
Concerns in Educational Research.....	34
Sample Size	34

Caliper width	35
Limitations in Covariate Selection	36
Chapter Three - Methods	39
Real Data Set	39
Participants	39
Software	42
Simulation Study	43
Research Questions	43
Initial Analysis	44
Research Question 1: Limitations in Sample Size	46
How does propensity score matching perform with limitations in sample size as measured by bias reduction?	46
Analysis	47
Research Question 2: Caliper Width	48
What is the ideal caliper width used for propensity score matching when a sample is limited in size, as determined by bias reduction?	48
Analysis	49
Research Question 3: Covariate Selection	50
What relationship of the covariates to treatment and outcome is optimal when sample size is limited, as determined by bias reduction and treatment effect?	50
Data Analysis Summary	52
Chapter Four – Results	54
Initial Analysis	54
Initial Match	58
Research Questions 1 and 2: Limitations in Sample Size and Caliper Width	66
How does propensity score matching perform with limitations in sample size as measured by bias reduction? What is the ideal caliper width used for propensity score matching when a sample is limited in size, as determined by bias reduction?	66
Research Question 3: Covariate Selection	72
What relationship of the covariates to treatment and outcome is optimal when sample size is limited, as determined by bias reduction?	72
Chapter Five – Discussion	82

Sample Size 82
Caliper Width 82
Covariate Selection 84
Limitations and Future Research 85

List of Charts

Table 1. Summary of Covariate Labels, Descriptions, and Codes	40
Table 2. Frequencies of Covariates.....	41
Table 3. Descriptive Statistics of GPA	41
Table 4. Q1: Limitations in Sample Size	47
Table 5. Q2: Caliper Width.....	49
Table 6. Q2: Caliper Width Conditions by Sample Size	49
Table 7. Q3. Covariate Selection Relations	52
Table 8. Correlations of Covariates to Treatment and Outcome	55
Table 9. Chi-Square Test of Covariates.....	56
Table 10. Independent Samples t-test for GPA before match	57
Table 11. Frequency of the Covariates before and after Match.....	59
Table 12. Summary of Balance for Matched Data	60
Table 13. Independent Samples t-test for GPA after match	65
Table 14. Summary of Balance for Matched Data (N=500).....	66
Table 15. Summary of Balance for Matched Data (N=400).....	67
Table 16. Summary of Balance for Matched Data (N=300).....	68
Table 17. Summary of Balance for Matched Data (N=200).....	68
Table 18. Mean c-statistic by Sample Size and Caliper Width.....	71
Table 19. Summary of Balance for Simulated Match (N=500).....	74
Table 20. Summary of Balance for Simulated Match (N=400).....	74
Table 21. Summary of Balance for Simulated Match (N=300).....	75
Table 22. Summary of Balance for Simulated Match (N=200).....	75
Table 23. Summary of Mean c-statistics by Sample Size for Simulation	78

List of Figures

Figure 1. Evidence Standards Decision Tree provided by the What Works Clearinghouse	2
Figure 2. Published Items Containing “Propensity Score” as the Topic by Year for Social Sciences.....	10
Figure 3. Citations containing “Propensity Score” by Year for Social Sciences.....	11
Figure 4. Published Items Containing “Propensity Score Matching” as the Topic by Year for Social Sciences.....	11
Figure 5. Citations Containing “Propensity Score Matching” as the Topic by Year for Social Sciences.....	12
Figure 6. General Procedure for Propensity Score Analysis	12
Figure 7. Jitter Plot of Original Data (N=1762).....	61
Figure 8. Histogram of the Original Data (N=1762)	62
Figure 9. Distribution of Final GPA Prior to Match.....	63
Figure 10. Distribution of Final GPA Prior for Control	63
Figure 11. Distribution of Final GPA Prior Treated.....	64
Figure 12. Distribution of Final GPA after Match.....	64
Figure 13. Bias Reduction by Sample Size and Caliper Width	69
Figure 14. Match Success by Sample Size and Caliper Width.....	70
Figure 15. Exclusions by Sample Size and Caliper Width	70
Figure 16. Mean c-statistic by Sample Size and Caliper Widths.....	72
Figure 17. Bias Reduction by Sample Size for Simulation	77
Figure 18. Summary of Mean c-statistics by Sample Size for Simulation	78
Figure 19. Bias in Estimated Treatment Effect for Matched and Unmatched.....	80

This page intentionally left blank

Chapter One - Introduction

Research in education has long been stigmatized as a pseudoscience due to the lack of random assignment and a true control group. A pseudoscience is defined as “a belief system, practice, or research that is presented as a science; however, it is not based on the scientific method and/or lacks scientific rigor, plausibility, or supporting evidence (Schutt, 2011).” In a true experiment, the random assignment of subjects to different groups guarantees that on average there should be no systematic differences in observed or unobserved covariates between those assigned to the treatment or to the control group. As a result, the treatment effect can be estimated by directly comparing outcomes between the treatment and the control group. Non-randomized, observational studies occur frequently in educational research in which there is no control over the treatment assignment; therefore, direct comparisons of outcomes from the treatment groups may be misleading. More specifically, evaluation research in education is typically limited to non-randomized, observational design yet has the additional task of quantifying the impact of an intervention.

Evaluation research started in the mid-1960s to evaluate large-scale government-issued programs with the purpose of improving the well-being of a specific sample of society (Barnow, Cain, & Goldberger, 1980). A journal named *Evaluation* was established in 1973 by a grant from the National Institute of Mental Health and the Evaluation Research Society was initiated in 1976 (Barnow et al., 1980). Evaluations were more of a cost-benefit analysis at this time, but quickly they have become a political focus to quantify the successes or failures of funded programs. Due to this, there has been a push in education for evidence-based practice accompanied by an increasing demand for evidence of efficacy of educational programs and interventions.

The What Works Clearinghouse (WWC) was developed in 2002 from a Department of Education initiative to inform educational research decision making in order to improve practice and the ability to demonstrate causal evidence. The What Works Clearinghouse reviews research studies and disseminates summary information and reports on over 6,000 publications on the WWC website. The WWC reviews each study and screens them to determine whether they provides strong evidence (*Meets Evidence Standards*), weaker evidence (*Meets Evidence Standards with Reservations*), or insufficient evidence (*Does Not Meet Evidence Standards*). The evidence rating relates to the amount of confidence one should place on the ability of the study to demonstrate causal evidence of the effectiveness of an intervention. These evidence standards provide researchers, educators, and policymakers a framework for how to create effective, inferential research designs. Figure 1 depicts the screening process for evidence standards and is provided in the WWC Procedures and Standards Handbook version 2.1.

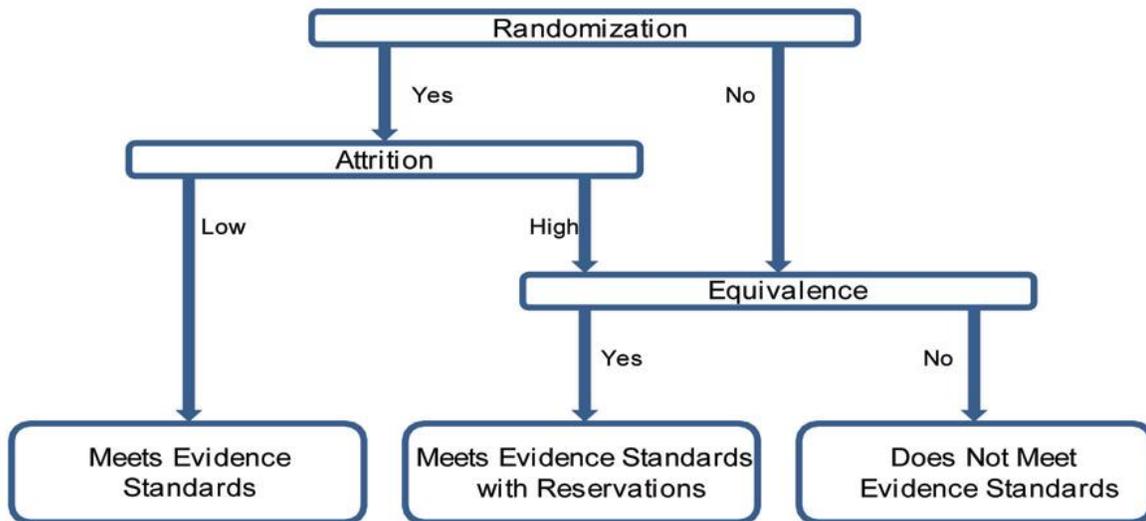


Figure 1. Evidence Standards Decision Tree provided by the What Works Clearinghouse

The evidence standards created by the What Works Clearinghouse require the randomization of subjects, equivalence of comparison groups, and accounting for attrition. Currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered *strong evidence (Meets Evidence Standards)*, while quasi-experimental designs (QEDs) may only be identified as *Meets Standards with Reservations*. In quasi-experimental designs, large differences on observed covariates in the two groups may exist, and those differences could lead to biased estimates of treatment effects. Because the groups may differ, a quasi-experimental design must be able to demonstrate that the intervention and comparison groups are equivalent on observable characteristics in order to meet evidence standards for the What Works Clearinghouse.

Recently, researchers have increasingly used methods based on propensity scores in order to account for the differences in baseline characteristics between the treated and control subjects in studies with quasi-experimental design. More specifically, researchers are using propensity score matching which involves selecting subsets of the treatment and control groups with similar covariate distributions (propensity scores) in order to control the confounding effects of these covariates that can create bias in the estimated treatment effects. The U.S. Department of Education supports propensity score matching as a method of evidence-based research when group equivalence can be established in the analysis (Lane, To, Shelley, & Henson, 2012). Therefore, the use of propensity score matching in educational research has steadily increased over the years. Recently, however, the utility of this method with small sample size has been argued. The purpose of this study is to analyze the efficacy of propensity score matching with the typical constraint of limited sample size in educational research.

Chapter Two - Review of Literature

The foundational purpose of most research is the study of cause and effect in order to duplicate or to prevent an event. Underwood defined the Principle of Causality in 1957 which states that “every natural phenomenon is assumed to have a cause, and if that causal situation could be exactly reinstated, the event would be duplicated (Maxwell & Delaney, 2004, p. 7).” Lazarsfeld described three criteria for causality in 1959 as 1) The causal relationship must have temporal order, in which the cause must precede the effect, 2) The two variables should be empirically correlated with one another, and 3) This observed correlation cannot be explained by a third variable (Guo & Fraser, 2014). However, there are many different variables that can generate a result, and it is impossible to determine all of those variables and the relationships among them. The acronym INUS, created by Mackie in 1974, most accurately describes the term ‘cause’ as an “insufficient but nonredundant part of an unnecessary but sufficient condition (Shadish, Cook, & Campbell, 2002, p. 4).” Thus, true effects are difficult to determine due to potential confounds.

The basis for statistical exploration of causal research is called the Neyman-Rubin Counterfactual Framework for causality (Neyman & Iwaszkiewicz, 1935; Rubin, 1974). This framework has also been called Rubin’s causal model and the potential outcomes model. It states that an inference about the impact of a treatment on an individual also involves a speculation of how the individual would have performed in the absence of the treatment. A counterfactual is defined as contrary to fact; it is the potential outcome that would have happened in the absence of the treatment or cause (Rubin, 1974; Shadish et al., 2002). Thus, a counterfactual is a missing value and is not observed. This framework emphasizes that there are potential outcomes for both

the treatment and control groups. In order to analyze these unobserved counterfactuals, research can be conducted by averaging the outcomes in other nontreated groups.

Experimental research is conducted in order to most accurately predict cause and effect. Experimental research studies are those that “deliberately vary something so as to discover what happens to something else later – to discover the effects of presumed causes (Shadish et al., 2002, p. 3).” A randomized control trial, or true experiment, has been deemed the gold standard in experimental research. In a true experiment, the random assignment of subjects to different groups guarantees that on average there should be no systematic differences in observed or unobserved covariates between those assigned to the treatment or the control group. Thus, the treatment effect can be estimated by directly comparing outcomes between the treatment and the control group due to random assignment. However, random assignment is not always possible, practical, or even ethical in the medical, behavioral, and social sciences. Non-randomized, observational studies occur frequently in educational research in which there is no control over the treatment assignment; therefore, direct comparisons of outcomes from the treatment and control groups may be misleading. These observational studies without random assignment are called quasi-experimental designs.

Since the 1940s, many fields, such as medicine and epidemiology, sociology, economics, education, and political science, have developed methods of estimating causal effects from observational data in order to overcome the stigma of quasi-experimental design, including matching. Matching has been defined broadly as “any method that aims to equate or balance the distribution of covariates in the treated and control groups (Stuart, 2010, p. 2).” Treated subjects are matched to similar nontreated subjects’ covariates with the intent of reducing the bias in

estimating the effect of the treatment. Covariates are defined as “a variable that is measured prior to the start of the treatment, such as age or gender, and hence is unaffected by the treatment (Joffe & Rosenbaum, 1999, p. 327).” Random assignment of subjects in experimental studies balances or controls for the unobserved covariate differences in groups. When random assignment is not possible, matching can balance the observed covariates.

Propensity Score Analysis

The use of propensity score analysis (PSA) was introduced by Heckman (1979) and Rosenbaum and Rubin (1983). Although both Heckman and Rosenbaum and Rubin discussed estimating treatment effects when the assignment of treatment was nonrandom, Heckman’s work used different terminology and focused mostly on the issue of sample selection. Therefore, Rosenbaum and Rubin are deemed the ones that first published this statistical technique. Propensity score analysis is a family of statistical techniques that utilizes propensity scores for causal inference when randomization is not feasible and includes matching, stratification, and weighting. Rosenbaum and Rubin define the propensity score as “the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum & Rubin, 1984, p. 516).”

The formula for calculating propensity scores is given below, where $e(x)$ is the abbreviation for propensity score, P a probability, $Z=1$ a treatment indicator with values 0 for control and 1 for treatment, the “|” symbol stands for conditional on, and X is a set of observed covariates.

$$e(x) = P(Z=1 | X)$$

In other words, the propensity score is the likelihood that a person would have been treated using only their covariate scores. Therefore, those subjects that have the same propensity score will have a similar distribution of covariates. Like all probabilities, a propensity score ranges from zero to one. The propensity score (also called the coarsest score) is a balancing measure because the treatment will be conditionally independent of the covariates for those subjects that have the same propensity score. Further, propensity score matching collapses and summarizes all observed covariates (also called the finest scores) into one scalar, the probability of being treated; more simply Rubin states, “the collection of predictors is collapsed into a single predictor (1997, p. 759).” The basis for using propensity scores relies on three theories developed by Rosenbaum and Rubin (1983, 1984). These theories are 1) Propensity scores balance observed covariates; 2) If it suffices to adjust for covariates, then it suffices to adjust for their propensity score; and 3) Estimated propensity scores are better at removing biases than true propensity scores because estimated propensity scores also remove chance imbalances on the covariates.

Assumptions

The Ignorable Treatment Assignment Assumption

Many sources of error can contribute to the interpretation of the outcome in observational research. Rosenbaum and Rubin (1983) named the fundamental assumption of propensity score analysis the ignorable treatment assumption. This assumption states that “conditional on the covariates X , the assignment of study subjects to binary treatment conditions (i.e., treatment or control) is independent of the outcome of nontreatment and the outcome of treatment (Guo & Fraser, 2014, p. 29).” Strong ignorability implies “that no systematic, unobserved, pretreatment

differences exist between treated and control subjects that are related to the response under study (Joffe & Rosenbaum, 1999, p. 329).” The ignorability assumption has been called many different things in the literature, such as unconfoundedness (Rosenbaum & Rubin, 1983), selection on observables (Barnow et al., 1980), conditional independence (Lechner, 1999), exogeneity (Imbens & Wooldridge, 2008), and common support. From this limited literature review, these terms are used interchangeably to denote the assumption that the outcome of either treatment or control is not conditional on placement in either group, as long as the covariates remain constant. The ignorability assumption is also the same as the assumption of ordinary least squares (OLS) regression, also called contemporaneous independence of the error term.

In quasi-experimental designs, this assumption is often violated due to group assignment being tied to outcome. To assess whether this assumption has been violated in observational studies, a chi-square test can be conducted when X is categorical and a t-test when X is a continuous variable. When the null hypothesis is rejected, stating that there are significant differences between the treated and non-treated groups, it may be concluded that this assumption has been violated and the treatment outcome is conditional on one of the covariates. Violating the ignorability assumption leads to biased and inefficient analyses of the treatment effect. Although these methods of testing the ignorability assumption are popular, Rosenbaum (2002) stated that this assumption is untestable due to the fact that no statistical evidence exists that supports the validity of these methods. Researchers can only build a convincing case that all important covariates have been included in the design. The most common approach then, is to make a diligent attempt to research likely covariates and then to include as many as possible.

The Stable Unit Treatment Value Assumption

Another assumption of propensity score analysis presented by Rubin (1980) is the stable unit treatment value assumption (SUTVA). This assumption implies that the outcome of the treated will be the same regardless of what mechanism was used to assign treatment. Further, SUTVA assumes that the value of the outcome for the treated individual will remain stable, regardless of other individuals receiving different treatments (Guo & Fraser, 2014). This assumption imposes what Heckman (2005) calls the following “exclusive restrictions”:

1) SUTVA rules out social interactions or general equilibrium effects, and 2) SUTVA rules out any effect of the assignment mechanism on potential outcomes (Guo & Fraser, 2014, p. 33).

According to Rubin, violations of SUTVA occur when the treated outcome is dependent upon which version of the treatment is received or when carryover effects occur in the treated and control groups that will impact the outcome.

Researchers can utilize propensity scores to balance non-equivalent groups using matching (Rosenbaum & Rubin, 1983), stratification (also called subclassification) (Rosenbaum & Rubin, 1984), or weighting (Hirano & Imbens, 2001) on the propensity score. Each of these techniques is an attempt to balance covariates prior to (matching and stratification) or while (stratification and adjustment) estimating the treatment effect. Research has demonstrated that propensity score matching can result in a greater reduction in bias than stratification on the propensity score (Austin, Grootendorst, & Anderson, 2007; Austin & Mamdani, 2006).

Propensity score matching is the more efficient technique and is most relevant to the field of education; therefore, it will be the focus of this study. Propensity score matching involves selecting subsets of the treatment and control groups with similar covariate distributions

(propensity scores) and matching them to estimate the causal effects of the treatment. Once matched on propensity scores, any differences between groups are thought to be estimates of the treatment effect.

Propensity score matching has been utilized more frequently in other fields, such as medicine, since its creation, but it has only recently gained popularity in the field of educational research. A search of the Web of Science was conducted to assess the frequency of published articles and citations from 1983 to 2014 using the words “propensity score” as the topic and found the total number of publications to date across all varying fields was 21,076. Of those publications, 5,868 were in the social sciences and were cited 84,487 times (Web of Science, 2105). The topic “propensity score matching” found 10,499 publications across all fields, with 2,907 from the social sciences with 40,009 citations. The figures 2 through 5 below provide a visual of the increasing trend in both publications and citations of propensity score analysis and matching in the social sciences after the initial publication by Rosenbaum and Rubin in 1980.

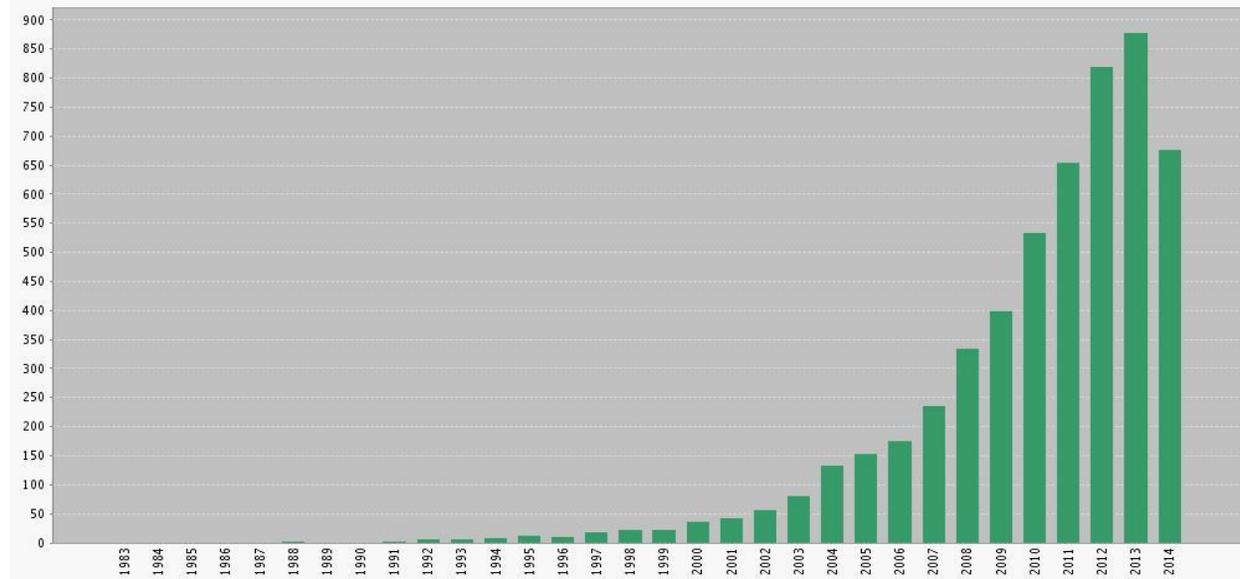


Figure 2. Published Items Containing “Propensity Score” as the Topic by Year for Social Sciences

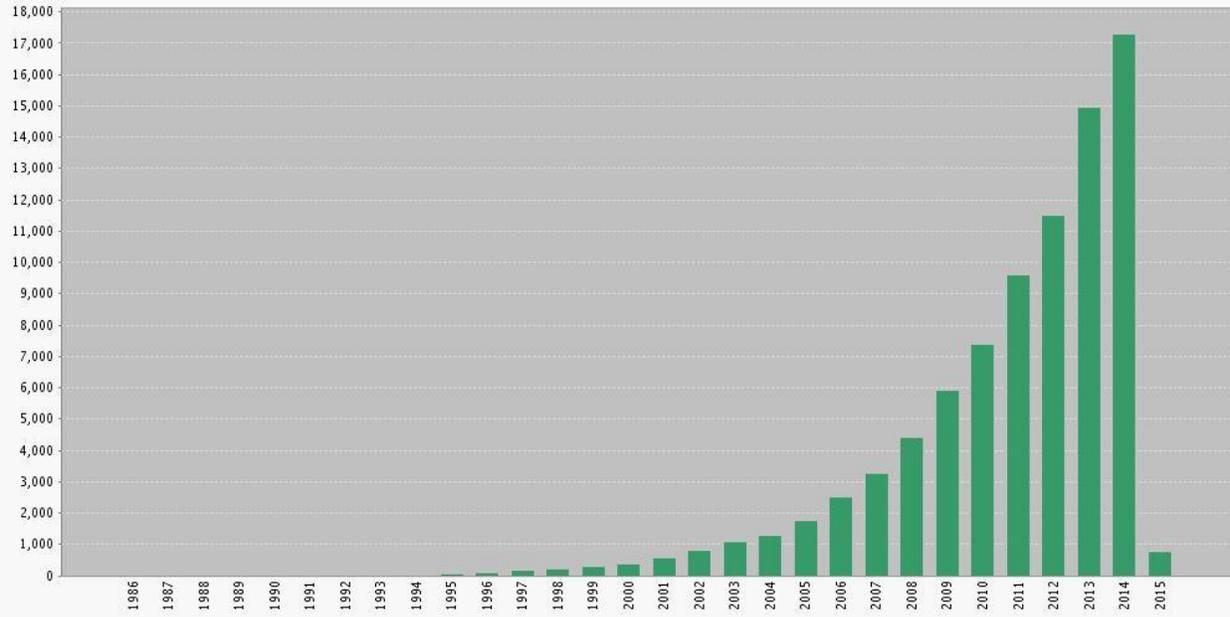


Figure 3. Citations containing “Propensity Score” by Year for Social Sciences

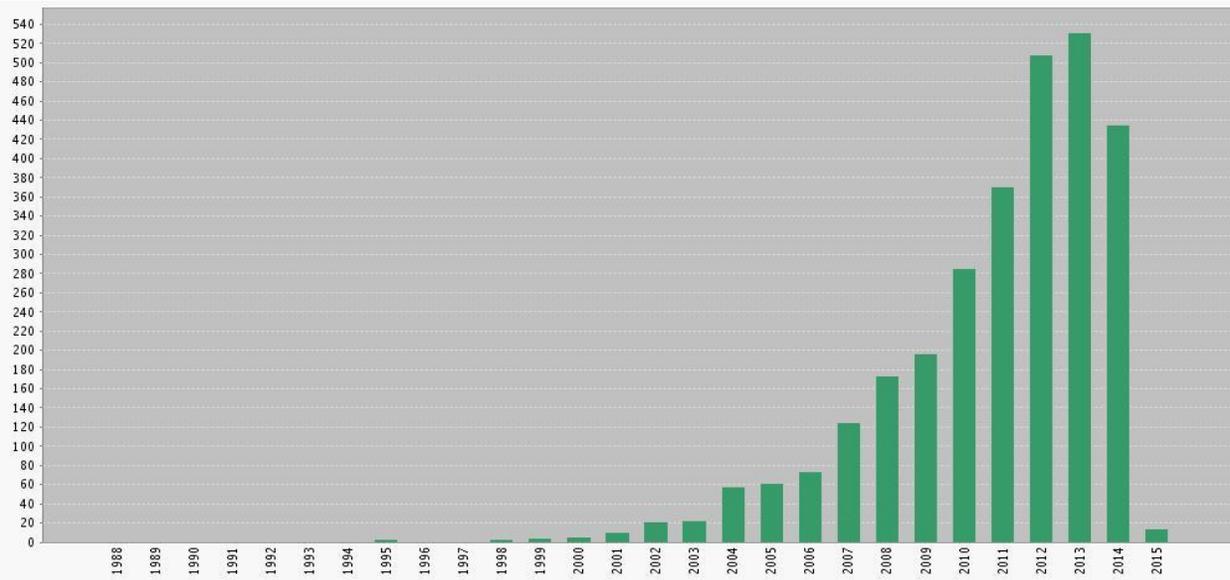


Figure 4. Published Items Containing “Propensity Score Matching” as the Topic by Year for Social Sciences

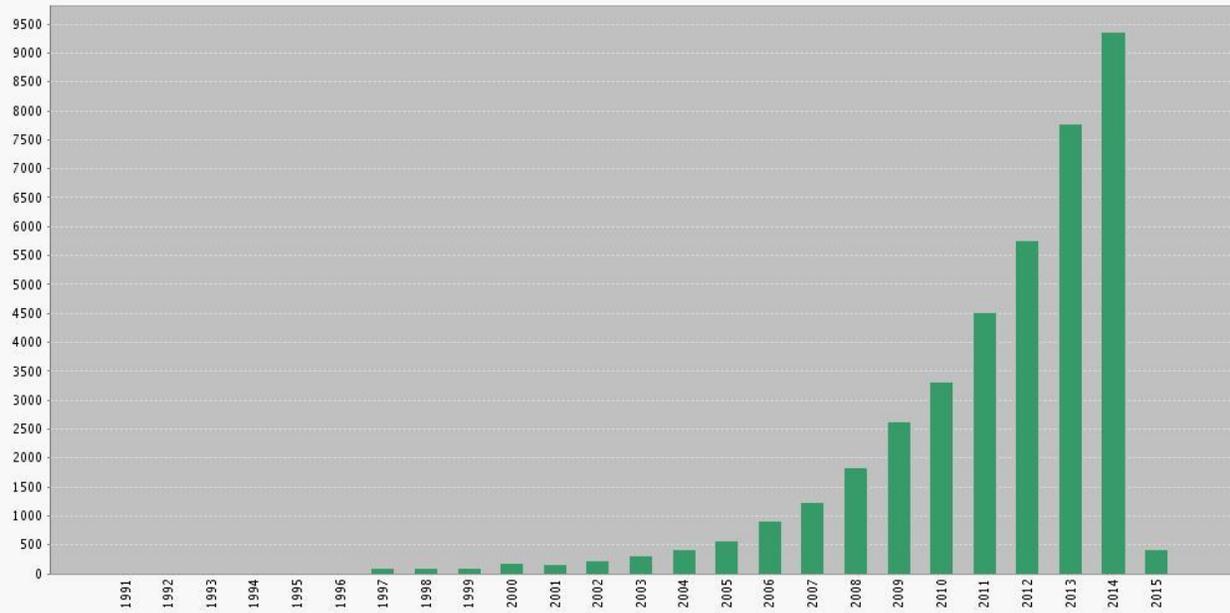


Figure 5. Citations Containing “Propensity Score Matching” as the Topic by Year for Social Sciences

These figures display the steady increase in publications and citations for propensity score analysis over time in the social sciences.

Steps for Propensity Score Matching

The figure provided below displays the steps specific to propensity score matching in propensity score analysis and will be discussed further.

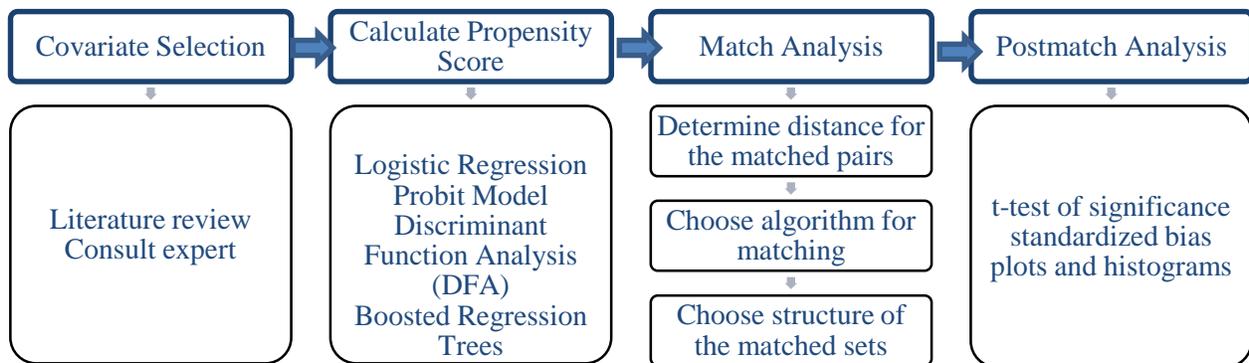


Figure 6. General Procedure for Propensity Score Analysis

Covariate Selection

The first step, and initial challenge, when conducting propensity score matching is identifying and selecting the potentially-relevant covariates. In a true experiment with random assignment, the definition of what constitutes a covariate is clear-cut; a covariate is any variable that can be measured before the assignment of treatment. Determination of what variables to include as a covariate becomes more challenging after treatment has been assigned. Some examples of possible covariates are age, ethnicity, years of education, years of experience, years of dosage, or marital status. Many studies have shown that the covariates selected can have a substantial impact on the performance of the propensity score (Heckman, Ichimura, Smith, & Todd, 1998; Lechner, 1999; Smith & Todd, 2005).

Several statistical techniques have been developed to decide which covariates to include in a propensity score analysis (Black & Smith, 2004; Heckman et al., 1998; Smith & Todd, 2005). These techniques, the hit or miss method, statistical significance, and leave-one-out cross-validation, are discussed briefly here. In the hit or miss method (also called trimming), only those variables that maximize the within-sample correct prediction rates are chosen. A covariate is classified as a “1” if the estimated propensity score is greater than the treated sample. If the covariate has an estimated propensity score that is less than the treated sample, it is a “0”. A similar method called minima and maxima comparison excludes all covariates whose propensity score is smaller than the minimum or larger than the maximum in the opposite group from the analysis. The statistical significance method is very common and relies on the statistical significance of a covariate. This method chooses a covariate that is a constant for the model (e.g. age) and then iteratively adds covariates assessing statistical significance. A covariate is kept if it

is statistically significant at the conventional or chosen level of significance. Heckman and his colleagues (1998) found that the prediction rates increase by a substantial amount when the hit or miss and statistical significance techniques are combined. The leave-one-out cross-validation procedure developed by Black and Smith (2004) is implemented by starting with a minimal model with only two covariates. Covariates are then added to the model while comparing the resulting mean squared errors. This method of covariate determination uses goodness-of-fit considerations instead of theory or evidence of the relationship of the covariate to the participation and/or outcome.

Often researchers assess the potential relevance of a covariate based on the statistical significant difference between the treatment and control groups. However, Rubin and Thomas (1996) suggest including all variables thought to be related to the outcome, regardless of whether they are related to the treatment. Specifically, they recommend that “unless a variable can be excluded because there is a consensus that it is unrelated to outcome or is not a proper covariate, it is advisable to include it in the propensity score model even if it is not statistically significant (1996, p. 253).” Including these seemingly unrelated covariates removes the nonsystematic bias due to the chance association between the covariate and exposure (Brookhart et al., 2006). Rosenbaum cautions against excluding covariates due to a lack of statistical significant difference in the treatment and control groups for the following reasons: 1) This analysis does not account for the relationship between the covariate and the outcome, 2) Statistical significance relies heavily on sample size and is not a prerequisite for practical relevance, and 3) This analysis would look at covariates individually, whereas the overall analysis will consider them collectively. Including all possible covariates known to be related to both the treatment and the outcome satisfies the ignorability assumption, which assumes that the assignment of subjects to

either treatment or control is independent of the outcome of treatment and control or that there are no systematic, unobserved, pretreatment differences between the treated and the control groups that are conditional on the unobserved covariates (Glazerman, Levy, & Myers, 2003; Guo & Fraser, 2014; Joffe & Rosenbaum, 1999; Rubin & Thomas, 1996). A complete and rich set of covariates is needed to satisfy the ignorability assumption; however, variables that may have been affected by the treatment should not be included in the matching process (Frangakis & Rubin, 2002; Greenland, 2003; Rosenbaum & Rubin, 1984).

Currently, there is no agreed-upon procedure or test available to provide guidance for researchers to know which covariates to include or exclude in a propensity score analysis. Ideally, knowledge of the subject matter and treatment would provide information regarding which covariates to select. Consulting with subject matter experts and conducting a literature review and pilot study to identify the relevant covariates is recommended (Luellen, Shadish, & Clark, 2005).

Calculate the Propensity Score

In observational studies the propensity score is unknown and can be estimated using any model that produces estimates of probability of group membership, such as logistic regression, the probit model, or discriminant function analysis. More recently, McCaffrey and colleagues employed methods based on boosted regression trees; however, this technique has not been widely used (McCaffrey, Ridgeway, & Morral, 2004). Logistic regression is more flexible than the other techniques and, unlike discriminant analysis, the predictors do not have to be normally distributed, linearly related to the dependent variable, or have equal variance within each group. Therefore, logistic regression is the prevailing approach.

Logistic regression allows one to predict a discrete outcome (e.g. group membership to the treatment condition in this case) from a set of variables that may be continuous, discrete, dichotomous, or a mix (Tabachnick & Fidell, 2013). The formula for using logistic regression is below where treatment status is regressed on covariates ($T_i=0/1$). In other words, the treatment assignment is the outcome variable predicted by the covariates. As a result, the collection of covariates is collapsed into a probability (propensity score) to having received the treatment.

$$\ln\left(\frac{P_i(T_i = 1)}{1 - P_i(T_i = 1)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_n X_{ni} + e_i$$

T_i = group (0 = control, 1 = treatment)
 X_{1i}, \dots, X_{ni} = scores on covariates
 e_i = random error

This equation creates the log of the odds, the probability of being in one group divided by the probability of being in the other group. Rosenbaum and Rubin (1985a) suggest using the logit of the predicted probability of the propensity score because the distribution approximates to normal.

The procedure for estimating coefficients for covariates in logistic regression involves maximum likelihood estimation. Maximum likelihood estimation is an iterative procedure that tries on arbitrary values of coefficients for the set of covariates and determines the direction and size of the coefficient necessary to maximize the likelihood of obtaining the observed frequency (Tabachnick & Fidell, 2013). The coefficients for the covariates, B , are the natural logs of the odds ratios. The odds ratio is the change in odds of being in one of the categories of outcome when the value of a covariate increases by one unit. Similar to linear regression, the coefficients are interpreted in relation to other covariates.

Goodness of fit statistic, chi-square χ^2 , is used to evaluate the covariates included in the model. In logistic regression, all covariates enter the equation simultaneously; thus, no hypothesis is made regarding order or weight of the covariates. An evaluation of the contribution of each covariate to the outcome should be assessed. A limitation of logistic regression is the interpretation of the results when covariates are correlated, as they often are. A covariate that is highly correlated with the outcome may show little predictive power in the presence of the other covariates. The chi-square statistic allows an analysis to be conducted to evaluate the impact of a covariate by comparing how the log-likelihood decreases or increases as covariates are added or deleted.

Tests of multicollinearity, tests of influential observations, and sensitivity analyses should be used to assess the fit of the model to the data. A number of statistics have been created to assess the goodness of fit of the model (e.g. Pearson chi-square, Hosmer-Lemeshow test, and pseudo R^2). The best logistic regression model is one that most closely estimates the propensity score to the participant's true propensity score. However, a participant's true propensity score is unknown; goodness of fit indices are the best indicators of successful prediction of propensity scores.

Matching Analysis

Although the matching process is conceptually straightforward, the method can be implemented using several different, complex procedures. Selecting which matching method to utilize depends on the ratio of treated to control subjects in the sample. For example, when the pool of control units is smaller than the treated units, several treated units might have to be matched with control subjects that are very different. Three issues arise when matching subjects

with controls: 1) Determining the distance for matched pairs, 2) Choosing an algorithm for matching, and 3) The structure of the matched sets. Each matching technique is a tradeoff between bias and variance. Which technique to choose is somewhat subjective, but is imperative to the results of any study using propensity score analysis.

Determining the distance for matched pairs

The simplest matching distance is called exact matching, which requires that the two groups be identical on the propensity score in order to match. However, exact one-to-one matching is difficult in practice; therefore, matching subjects with comparison units whose propensity scores are sufficiently close typically becomes the objective. This alternative to exact matching is called approximate matching. In approximate matching two units' propensity score can be matched if their propensity scores are approximately the same based on the Mahalanobis Distance Measure (MDM) or caliper.

Mahalanobis Distance Measure

Propensity score matching has been combined with various other matching techniques to increase robustness and decrease bias. The most common combination is propensity score matching with the Mahalanobis distance measure technique (Rosenbaum & Rubin, 1985b; Rubin, 1980). The Mahalanobis distance measure (MDM) was created by P.C. Mahalanobis in 1936 Mahalanobis (1936). This technique randomly orders subjects based on several background variables and then calculates the distance between the first treated subject and all control subjects. The formula for calculating the Mahalanobis distance measure is provided below where u and v are values of the covariates for the treated subject i and control subject j and C is the

sample covariance matrix of the matching variables for the full set of control subjects (Guo & Fraser, 2014).

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

The control subject, j , with the smallest distance is matched to the treated subject, i , and then both subjects are removed from the pool. This process is repeated until matches are complete for all treated subjects. One of the drawbacks of this technique is that as the number of covariates increases, so does the distance between the treated and control subjects, making close matches difficult (d'Agostino, 1998). When using the Mahalanobis metric matching with propensity scores, the propensity score is an additional covariate.

Matching within calipers

A caliper is a preset tolerance for the distance between the propensity scores of the treated and control subjects to enable the matching (Rosenbaum & Rubin, 1984). Using calipers overcomes the subjectivity of erroneously choosing a control participant to match to a treated participant. The formula used to determine a match based on calipers is below, where ε is the previously specified caliper.

$$\|P_i - P_j\| < \varepsilon$$

All control subjects within a caliper of the treated subject's estimated propensity score (or estimated logit of the propensity score) are selected, and the closest control subject and the treated subject are then matched and removed from the pool. The process is then repeated. All remaining control subjects are available for the next matching with a treated subject. The size of the caliper is determined a priori by the investigator, because it is difficult to know what

tolerance level is reasonable. Cochran and Rubin (1973) and Rosenbaum and Rubin (1985b) suggest that caliper size of a quarter of a standard deviation of the logit of the propensity score be used to reduce 90% of the bias in propensity scores.

Choosing an algorithm for matching

This step involves matching treated to control subjects based on their propensity scores. Matching approaches are used when the goal is to reduce as many possible differences between groups as possible; consequently, not all subjects are retained. Matching is often referred to as resampling due to the loss of subjects without a good match. Different matching techniques exist to try to retain as many subjects as possible.

Nearest available matching

Nearest available matching, also called nearest neighbor or greedy matching, is the most common matching algorithm and consists of placing the treated subjects in random order and then selecting the nearest control subject with the closest propensity score (Rubin, 1976). Both subjects are then removed from consideration for matching and the next treated subject is selected, repeating this process for all unmatched treatment subjects until all are matched. If more than one control subject is needed for each treated subject, then once every treated subject has one control the first treated subject is assigned to the nearest of the remaining available controls. Nearest available matching is the most prevalent matching technique, but it does not minimize the total distance within matched pairs (Gu & Rosenbaum, 1993).

Optimal Matching

An alternative to the nearest available technique is optimal matching. This matching method can minimize the total within-pair difference of the propensity score by finding the matched pairs with the smallest average distance across all pairs. Optimal matching is most appropriate when there are not many control matches for the treated subjects. The two options (nearest available and optimal) were found to be comparable in terms of producing a balanced matched sample, but optimal matching was found to be better at minimizing the distance within the pairs (Gu & Rosenbaum, 1993). This matching technique is more complicated and time-consuming, yet it is slowly increasing in use and feasibility due to the availability of software programs and packages that can perform optimal matching quickly and efficiently.

The structure of the matched set

After the algorithm for matching is determined by the researcher, the structure of the matched set must also be decided. In matching with replacement (also called one-to-many), a treated subject can be matched with more than one control subject, depending on the availability of adequate matches. In other words, the treated subject is matched with the control subject regardless of whether that control subject has been previously matched. This matching technique minimizes the propensity score distance between the matched control and the treated subject and is beneficial in bias reduction (Dehejia & Wahba, 2002). Matching with replacement becomes ideal when the distributions of the estimated propensity scores (region of common support) are very different. For example, a sample may have a treated group with high propensity scores and very few control subjects with high propensity scores to match with. In this example, allowing

replacement will reduce the number of control subjects used to construct the matches and will increase the variance (Smith & Todd, 2005).

A special case of a one-to-many match (with replacement) is called full matching, where each matched set can contain one treated unit with one or more control or one control matched with one or more treated subjects (Gu & Rosenbaum, 1993). Matches are created based on the closest propensity score distance between any treated subject and control subject, not based on a previously determined distance or number. Gu and Rosenbaum conducted a Monte Carlo simulation and found that full matching is better than one-to-many matching; forcing every treated unit to have k control matches created poor matches (1993).

In matching without replacement (also called one-to-one matching), a control subject is no longer available to match for subsequent treated subjects once the control subject has been matched to a given treated subject. In order to ensure the smallest possible propensity score distance between the treated and control units, matching without replacement would be used where a single control unit is utilized for a single treated unit.

When there are fewer comparison subjects similar to the treated subjects, forced matching may occur on propensity scores that are quite different, increasing bias (Dehejia & Wahba, 2002); however, using more control units per treated unit in matching with replacement increases the precision of the estimates while also increasing bias. An additional concern for matching without replacement is that the results can be sensitive to the order in which they are matched (Rosenbaum & Rubin, 1985a). In practice, the selection of a method becomes difficult and is dependent on both the subjective decision of the researcher and the specific dataset.

There are many different types of matching techniques and each researcher must decide which matching algorithm is right for their particular dataset, the treated and control groups' common region of support aids in this decision. The common support region is the overlap between the distributions of the propensity scores of the treated and the control groups. The distributions of the propensity scores can be assessed visually by looking at a Q-Q plot or a jitter plot. The common support region informs the researcher about sufficient overlap between the two groups and is linked to the type of generalization of the causal effect that can be made. A broad region of common support allows causal effect estimates over the full range of the propensity score in the sample, whereas small common support regions restrict the estimation of a causal effect to a subsample. Further, observing the common support region ensures that there are common characteristics in both the treated and control groups and is further evidence that matching is a possible analysis for the sample (Bryson, Dorsett, & Purdon, 2002). When there is substantial overlap in the propensity score distributions, the matching methods will yield similar results. When the distributions of the propensity scores are very different for the treated and control groups, finding a satisfactory match without replacement can be difficult. It has been advised by Imai and colleagues to exclude the units that fall outside the common support region, as no causal effect is defined for these units (Imai, King, & Stuart, 2008). However, Bryson and his colleagues (2002) caution that discarding a large number of observations causes a concern for the representativeness of the remaining sample. They recommend looking at the characteristics of the discarded observations for possible important distinctions when providing estimates of the treatment effects (Bryson et al., 2002).

Postmatch Analysis

Assessing the Matching Quality

In observational studies, researchers are concerned with threats to internal validity, or factors other than the intended intervention that could impact the evaluation of the treatment effect. Messick (1989) provided the most modern view of validity as how relevant the test scores are at measuring the construct. Validity is now seen as a unitary concept with construct validity being the integrating force in which there is only evidence of validity and includes the consequences of test use/misuse. The nine common threats to internal validity include: Attrition, regression to the mean, maturation, selection bias, history, instrumentation, testing effects, construct-underrepresentation and construct-irrelevant variance (Shadish et al., 2002). These threats to validity have been studied extensively and are collectively named “bias” when assessing the quality of matches in these analyses.

Rosenbaum (2002) discussed the two main types of bias found in observational studies: overt bias and hidden bias. Overt bias can be observed in the data, while hidden bias cannot be observed because the required information to reveal a bias was omitted from data collection. Overt bias can be assessed and accounted for, while hidden bias cannot be directly corrected for. Rosenbaum (2002) provides a method for assessing hidden bias in a sensitivity analysis. A sensitivity analysis examines whether the qualitative conclusions drawn would change in response to hypothetical hidden biases of varying magnitudes. However, it is important to recognize that propensity score matching only accounts for overt bias due to the inclusion of only observable covariates. Only randomized control trials (when conducted appropriately) can control for both overt and hidden bias.

Randomized experiments control for hidden bias and therefore remain the gold standard in practice. Selection bias occurs when the treatment or control status of subjects is related to unmeasured or unobserved characteristics that are related to the outcome in question. The term bias refers to the potential to misinterpret estimations of the treatment effect on the outcome (Barnow et al., 1980). Selection bias is a departure from the strongly ignorable treatment assignment assumption, which assumes that given balance in the covariates; there are no measured or unmeasured differences other than the treatment received (Rosenbaum, 2002). The propensity score matching method relaxes the ignorability assumption by resampling the treated and control groups by the covariates, so they become more similar to randomly assigned groups. Bias in observational studies exists and analyses should be conducted before generalizing findings to a population.

It is important to examine whether the distribution of covariates is similar between the treated and control subjects in the matched sample both before and after matching. Bias in the matched pairs is assessed by determining the average matched pair difference due to incomplete matching (the failure to match all treated units by discarding some treated units as unmatchable) or inexact matching (the failure to obtain exact matches) (Rubin, 2006). Rubin and Thomas (1996) provide approximations that can be used to determine the possible bias reduction from matching in a specific dataset based on the initial difference in the covariates between the groups, sample size, the number of matches desired, and the correlation between the covariates and the outcome. If considerable differences are found, Rosenbaum and Rubin (1985b) suggest including additional covariates, interactions between covariates, or polynomial terms to model the nonlinear relationships between treatment and covariates. Several procedures exist to analyze the balance of the covariates both before and after matching.

There are two different aspects for evaluating matches. They can be evaluated on distance and balance. The distance is the difference between the propensity score values of the matched treated subject and the control subject(s). The distance of the propensity scores prior to the match is determined using the Mahalanobis distance measure or calipers (discussed previously). If a matched set contains a treated subject to one or more control, then there are multiple distances to assess. For example, when a treated subject is matched to two control subjects, there is a distance between the treated and the first control and a second distance between the treated and the second control. The average distance within matched sets is the calculated average over all matched sets of all the distances within the matches (Gu & Rosenbaum, 1993). A small distance implies that the pairs are comparable in terms of covariates.

The balance is defined as “the similarity of the empirical distributions of the full set of covariates in the matched treated and control groups (Stuart, 2010, p. 11).” In other words, a match is considered balanced when the distributions of the covariates in the treated and control units being matched are similar. A balanced match implies that the matched pairs are comparable as a whole. There are several different methods used to evaluate the balance of the matched pairs; each of these will be discussed below.

t-test of significance

Balance can be calculated using a bivariate test (e.g. Wilcoxon rank sum test, an independent sample t test, or a one-way analysis of variance (ANOVA)) for a continuous covariate, or a chi-square test for a categorical covariate before and after matching and is similar to an effect size. If the bivariate test shows significant differences between the treated and control groups on a covariate prior to matching, then the covariate needs to be controlled for by

including it as a covariate in the propensity score analysis. If the postmatching bivariate test shows significant differences, the model used to predict the propensity score should be reevaluated. If the postmatching bivariate test is nonsignificant, the conclusion is that the propensity score has successfully removed group differences based on the covariates. Matches are considered to be balanced when the differences in the means of the covariates are less than .25 and the variance ratios are between .5 and 2 (Rubin, 2001).

Utilizing t-tests to address imbalance by observing the difference in means for each variable in the treated and control groups is misleading according to Imai, King, and Stuart (2008) and Austin, Grootendorst, and Anderson (2007). The t-test is a function of both balance and power. Finding matched pairs of treated subjects to control subjects ultimately demands larger numbers of control units being discarded, decreasing power; therefore, the value of the t-statistic becomes closer to zero, falsely indicating improvements in balance. Smaller sample sizes typically produce less power and will falsely inflate the p-value (Imai et al., 2008). Furthermore, Ho and colleagues state that the use of the hypothesis testing is inappropriate in this context as balance is a characteristic of a sample, not a hypothetical population (2007). Alternatively, researchers suggest evaluating balance using difference in the means first but then following up with higher-order evaluations such as non-parametric density plots, propensity score summary statistics, or a quantile-quantile (QQ) plot (Austin & Mamdani, 2006; Imai et al., 2008). They identified the two key features when evaluating balance as, 1) the statistic should be a characteristic of the sample and not of a hypothetical population and 2) the sample size should not affect the value of the statistic (Austin, 2008; Imai et al., 2008).

Standardized Bias

An alternative to assess the balance of the propensity score is by looking at the differences between the means of each covariate divided by the standard deviation in the total treated group, called ‘standardized bias’ or ‘standardized difference in the covariate means’ before and after matching (Haviland, Nagin, Rosenbaum, & Tremblay, 2008). This is also referred to as the percent reduction in bias or B , which is used widely when evaluating matching (Rubin, 1980). Haviland and colleagues developed the formula below to calculate the absolute standardized difference in the covariate means prior to matching, where M_{xt} and M_{xp} are the means of X for treated and potentially control groups, respectively.

$$d_x = \frac{|M_{xt} - M_{xp}|}{S_X}$$

After matching, the balance in the covariates is assessed with the following formula, where c denotes the control group and M_{xc} denotes unweighted mean of the means of the covariate X for the control matched to the treated.

$$d_{xm} = \frac{|M_{xt} - M_{xc}|}{S_X}$$

The absence of significant differences between the treated and control groups after the match is acceptable evidence that balance has been achieved. A concern with the standardized bias approach is that there is not a defined level of success of the matching procedure in the literature. In most empirical studies a bias reduction below ten percent is seen as sufficient (Austin et al., 2007; Austin & Mamdani, 2006; Stuart, Marcus, Horvitz-Lennon, Gibbons, & Normand, 2009), while others state the need for five percent or less (Caliendo & Kopeinig, 2008).

Plots and Histograms

Multidimensional histograms, a quantile-quantile (Q-Q) plot, or jitter plots can be used to observe the distributions of the covariates before and after matching. The Q-Q plot displays quantiles of a covariate in one group against quantiles of the same covariate in the other group. If the two distributions exhibit perfect distributional equivalence all data points in the Q-Q plot will fall on one single line, the 45 degree diagonal through the plot. Deviations from balance are characterized by data points that fall above or below this diagonal. The advantage of the Q-Q plot is the ease with which deviations can be detected.

The common support region that quantifies the amount of overlap on the propensity score between the two groups can easily be assessed by observing the range of the distributions, or graphically by overlaying the histograms of the propensity score in each group. The R program *twang* (Ridgeway, McCaffrey, Morral, Burgette, & Griffin, 2013) and *MatchIt* (Ho, Imai, King, & Stuart, 2004) produce graphical representations for checking imbalances in covariates both before and after matching.

Propensity Score Summary Statistics

The c-statistic is a measure of the predictive ability of a model and is commonly used to assess the quality of a propensity score. It calculates the proportion of pairs in which the treated subject had a higher estimated propensity score than the control subject (Harrell, 2001). It is used to compare the goodness of fit of the logistic regression model. Values for this measure range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of membership in a group, and a value of 1.0 indicates that the model perfectly

identifies those within a group and those not. Models are typically considered reasonable when the c-statistic is higher than 0.7 and strong when c exceeds 0.8 (Hosmer, Lemeshow, & Sturdivant, 2000). The c-statistic has also been used to identify which covariates to include or possibly exclude in a research design. Any covariate that increases the c-statistic, or the predictive ability of the model, would be included in the design (Brookhart et al., 2006); other researchers have argued that the c-statistic is not an accurate measure of classification ability because it does not use covariate balance as a criterion (Augurzky & Schmidt, 2001; Brookhart et al., 2006; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008). Including the c-statistic, when evaluating the overall balance, remains common practice.

Estimating the treatment effect

After the treated and control subjects are matching on propensity scores, any multivariate analysis that allows for matched pairs can be completed. However, most multivariate analyses are only permissible for matched samples created by nearest neighbor or greedy matching. Matches created by optimal matching require a different type of regression adjustment, such as the Hodges-Lehmann aligned rank test. Due to this additional requirement for analysis, most propensity score matching utilizes nearest neighbor or greedy matching.

The method chosen for analyzing the matched pairs for the treatment effect must account for the matching procedure. Propensity score matching is part of the design of the study and not part of the actual analysis; therefore, the actual analysis for the causal inference must include methods that account for the matched nature of the sample data. Regression-based methods or models can be utilized as long as the assumptions are met (Austin, 2008).

Advantages

Matching methods have a few advantages over other alternate approaches that control for background variables (regression, structural equation modeling, or selection models) such as highlighting insufficient overlap between the covariates in the treatment and control groups. Regression and selection models have been shown to perform poorly when there is not sufficient overlap, but these procedures do not involve checking for this overlap (Dehejia & Wahba, 2002; Glazerman et al., 2003). Matching methods assess the amount of overlap in the covariates, thus making the researcher aware of the quality of the resulting inferences that would be made. Further, matching methods have straightforward diagnostics and procedures by which the quality and performance can be assessed.

Like other matching procedures, propensity score matching estimates an average treatment effect from observational data. After Rosenbaum and Rubin (1983) designed propensity score matching, many studies regarding the effectiveness of propensity score matching have occurred (Glynn, Schneeweiss, & Stürmer, 2006; Heckman et al., 1998; Kurth et al., 2006). Prior to propensity score matching, traditional methods of adjustment (matching, stratification and covariance adjustment) were often limited because they could only use a limited number of covariates for adjustment. However, propensity scores, which provide a scalar summary of the covariate information, do not have this limitation and that is considered its' key advantage. If the treatment and control were balanced on covariates one at a time, large numbers of observations would be needed. In a simulation conducted by Gu and Rosenbaum (1993), they found that propensity score matching outperformed other matching techniques due to this ability

to balance on many covariates simultaneously, potentially approximating the balance similar to randomization.

In quasi-experimental studies the propensity scores are estimated due to the fact that the true propensity is unknown. One would expect that an estimated value would not perform as well as a true value, but researchers have found the opposite in theory, simulation, and in practice. The estimated propensity scores remove some of the chance imbalances in the propensity scores that the true propensity scores leave behind.

Propensity score matching can be especially useful in research studies in which there is a small number of subjects that received the intervention or treatment and a larger number of control subjects that did not. Contrary to the Mahalanobis matching technique (matching is done by randomly ordering subjects and then calculating the distance between the first treated subject and all controls and repeated) in which it is difficult to find close matches when there are many covariates, propensity score matching can be completed with many covariates and is typically easier to find matches (Gu & Rosenbaum, 1993). Propensity matching was found to remove more than twice the bias removed by Mahalanobis metric matching when a simulation study was conducted with 20 covariates (Gu & Rosenbaum, 1993; Joffe & Rosenbaum, 1999).

Limitations

Based on the literature, there are three limitations of propensity score matching (Guo, Barth, & Gibbons, 2006; Rosenbaum & Rubin, 1983; Rubin, 1997). One limitation of is that propensity scores can only be attained from observed (and observable) covariates. Factors that affect assignment to treatment but that cannot be observed cannot be accounted for in the

matching procedure. Therefore, the accuracy of estimates from propensity score matching could be seriously affected by missing predictors or confounds (Bai, 2011; Joffe & Rosenbaum, 1999; Weitzen, Lapane, Toledano, Hume, & Mor, 2004). Imai and colleagues discuss the possibility of error in observational studies due to the imbalance in unobserved variables, the “Achilles heel of observational studies (Imai et al., 2008, p. 493).” To deal with hidden bias, Rubin (1997) recommends conducting sensitivity analysis and testing different sets of conditioning variables to address this limitation. Secondly, propensity score matching requires large samples, with substantial overlap between treatment and control groups (Rubin, 1997; Weitzen et al., 2004). Smaller sample sizes have less overlap between treated and control groups, which could result in discarding more units due to lack of matches and ultimately even fewer matches for analysis (Weitzen et al., 2004). Finally, propensity scores consider the covariates that are related to the treatment but not the outcome the same as those that are related to the treatment and strongly related to outcome (Rubin, 1997). Inclusion of covariates that are only slightly predictive of the outcome reduces the efficiency of the relevant covariates. However, this is a feature that propensity score matching shares with randomization. Further, Rubin and Thomas discovered that the bias effects for not including a weak predictor covariate override the efficiency gains (1996). Judea Pearl has also argued that hidden bias may actually increase because matching on observed variables may unleash bias due to dormant unobserved confounders. Similarly, Pearl has argued that bias reduction can only be assured (asymptotically) by modeling the qualitative causal relationships between treatment, outcome, observed and unobserved covariates (2009, 2011).

Concerns in Educational Research

This literature review has provided an overview of the methodology necessary to utilize propensity score matching. Many studies have been conducted to demonstrate the effectiveness of propensity scores at increasing precision by reducing the variances of the distributional differences across treated and control groups in quasi-experimental designs. Recent work, however, has shown that the degree of bias reduction achieved and the efficiency of the covariates as estimators can differ greatly based on the methodology used to perform the matching as it relates to sample size, the caliper width, and the relation of the covariates to treatment or outcome (Althausen & Rubin, 1970; Angrist & Hahn, 2004; Frölich, 2004; Glazerman et al., 2003; Zhao, 2004).

Sample Size

The need for a large sample size when conducting propensity score matching, to allow for overlap between the treated and control groups with many potential relevant covariates has been documented (Dehejia & Wahba, 2002; Fan & Nowell, 2011; Glazerman et al., 2003; Lane et al., 2012; Luellen et al., 2005; Rubin, 1997; Smith & Todd, 2005; Yanovitzky, Zanutto, & Hornik, 2005). It seems logical that in order to have sufficient matches, there must be a large pool of potential matches in the nontreated group. Some research suggests that propensity score matching may only be feasible when there are extremely large sample sizes over 10,000 (Peikes, Moreno, & Orzol, 2008). However, an overly large sample of controls is usually impractical in educational research due to difficulty of obtaining consent, finding a comparable group, or cost.

For research in the social sciences it is typical to have non-random convenience samples, such as from one school, which they were able to gain access to or permission to study. These

research designs are fraught with small sample sizes in the treatment groups and possibly even smaller nontreated groups. A typical sample size in educational research could be less than 500. Literature on the efficacy of propensity score matching with sample sizes less than 500 is limited. Recent research on propensity score matching found that sample sizes less than 300 may be too small for matching when prediction of group assignment is high (Lane et al., 2012). Studies of propensity score matching and sample size agree that small sample sizes could lead to poor matches and unstable estimations of the treatment effect. Further, large sample sizes are required for propensity score analysis in order to balance the covariates across groups and reduce bias, but the required minimum for sample size is yet to be determined for propensity score matching.

Caliper width

The most common method used for propensity score matching is nearest neighbor (greedy) matching within fixed caliper widths. As mentioned previously, a caliper is a preset tolerance for the distance between the propensity scores of the treated and control subjects, which enables matching (Rosenbaum & Rubin, 1984). The size of the caliper is determined a priori by the investigator, but at what tolerance the caliper width should be set is still in question. Most researchers use a caliper width 0.2 standard deviations of the logit of the propensity score due to the research by Cochran and Rubin (1973) and Rosenbaum and Rubin (1985b) suggesting that this caliper size would reduce 99% of the bias in propensity scores (Austin, 2008, 2009; Austin et al., 2007) while others have used a caliper width of 0.6 as this same research stated this caliper width would reduce 90% of the bias (Ayanian, Landrum, Guadagnoli, & Gaccione, 2002). Propensity score matching has been used more frequently in medical research, where

caliper widths can vary greatly. This limited literature review found researchers that have used a caliper width of 0.005 (Christakis & Iwashyna, 2003; Cole et al., 2002), 0.03 (Yu et al., 2003), 0.02 (Murray, Singer, Dawson, Thomas, & Cebul, 2003), 0.01 (Hall, Summers, & Obenchain, 2003; Magee, Coombs, Peterson, & Mack, 2003; Seeger, Walker, Williams, Saperia, & Sacks, 2003), and 0.1 (Moss et al., 2003).

With limitations in sample size, the question of what caliper width to use becomes even more complicated. The choice of caliper width should be a decision of the variance-balance trade off. However, caliper width will impact the final sample size of the successful matches. A narrower caliper width will result in matching more similar subjects, reducing bias by reducing the systematic differences in the treated and nontreated groups, but will also reduce the final matched sample size and will increase the variance in the estimated treatment effect. Using a wider caliper width will retain more subjects, decrease the variance in the estimated treatment effect, but will increase the bias and systematic differences in the two groups. Currently, there is not been an identified optimal caliper width as it relates to studies with limitations in sample size.

Limitations in Covariate Selection

Rubin and Thomas (1996) used approximations for the reduction in bias and variance and suggested including all covariates thought to be related to the outcome regardless of their relationship to the treatment. Later, Rubin further supported this by stating that including variables that are strongly related to the treatment but unrelated to outcome will decrease the efficiency of the estimated treatment effect; however, he also stated that excluding those variables would actually cause more concern with estimating the treatment effect than leaving

them in the design when samples are substantial in size (1997). These results have been replicated by other studies as well (Brookhart et al., 2006; Drake, 1993; Perkins, Tu, Underhill, Zhou, & Murray, 2000; Robins, Mark, & Newey, 1992). Conversely, a Monte Carlo simulation conducted by Brookhart and his colleagues found that it would be advantageous to exclude covariates that are confounders in small studies ($n=500$) that are strongly related to the treatment and only weakly related to the outcome (2006). The inclusion of such a covariate and the increase in variance is not offset by a decrease in bias to improve the mean squared error. As sample size increased ($n=2500$), they found that the variance of the covariate decreases proportional to $1/n$, yet the bias remains. They concluded that it would not be recommended to exclude covariates related to treatment in moderately sized studies unless it was known to be unrelated to the outcome. However, these studies have not analyzed the inclusion or exclusion of covariates as it relates to treatment and/or outcome with sample sizes less than 500.

With limitations in sample size, comes a limitation in the number of possible covariates to include in the matching design. In small samples, it may not be possible to include a very large set of variables in the procedure. With limited covariates, Brookhart and colleagues (2006) suggest that the priority should be given to variables that are believed to be related to the outcome, as choosing variables with low relation to the outcome and high relation to the treatment will result in increased variance. Another strategy when working with a limited sample size, is to include a small number of covariates that are known to be related to the outcome, conduct the matching and then assess the variance on all of the available covariates, including any additional variables that remain unbalanced after the match (Stuart, 2010). Specific strategies of covariate selection are yet to be determined when working with limited sample sizes.

The propensity score matching technique has considerable potential for educational research given the likelihood for selection bias due to programs being implemented that are intended for special populations, the seldom use of random assignment, and the examination of only the observed covariates. Research with quasi-experimental design must be able to demonstrate that the intervention and comparison groups are equivalent on observable characteristics in order to meet evidence standards for the What Works Clearinghouse. The U.S. Department of Education currently supports the use of propensity score matching as a method of evidence-based research when group equivalence can be established in the analysis (Lane et al., 2012). However, further research must be done to ensure that causal inference and impact estimates are effectively reducing bias when using propensity score matching with limited sample size. The current research on propensity score matching is missing imperative information for educational researchers regarding the practical implications of utilizing this method with sample sizes smaller than 500. The objective of the current study is to assess the effectiveness of propensity score matching at reducing bias in a real data set with limited sample size as compared to an ideal data set created in a Monte Carlo simulation study. It is hypothesized that the effectiveness of propensity score matching at reducing bias will drop below acceptable levels for research when sample size is reduced to a certain point.

Chapter Three - Methods

The purpose of this study was to evaluate the effectiveness of propensity score matching in quasi-experimental research when limited by sample size as determined by a reduction in bias. In an effort to determine the optimal selection of covariates and caliper width with a limited sample size, this research included both simulated and real data. This chapter describes the participants, data cleaning procedures, software, experimental design and methods used for the real data study as well as the experimental design and methods for the simulation study. The assumptions, statistical analysis, and limitations of this study will also be discussed.

Real Data Set

Participants

Data for this study used a real life dataset acquired from a local school district that partnered with a federally-funded program aimed at helping minority students gain access to the skills necessary to pursue and complete post-secondary education. The intervention was implemented in the seventh grade and continued through high school graduation. This dataset was collected when the students were in the second year of the program. The original data file from the school district that was shared included 2059 subjects. Those subjects with any missing data were deleted, which resulted in 307 subjects being removed out of the original dataset or 14.9%. The resulting dataset included 1752 students within one school district; schools that participated and some that did not were included, deriving comparable samples of students with similar covariates that were in the treatment and the control group. Students that did not disclose their ethnicity status (N=26) were then removed from the dataset, resulting in a sample size of 1726 students (886 having received the intervention and 840 that had not received the

intervention). I received access to this data with a data sharing agreement with the Kansas City, Kansas Public School District (See Appendix A). The data contained no identifiable information. Human subject approval was received from the Institutional Review Board at the University of Kansas (See Appendix B).

The matching variables provided from the district and how they were coded is described in Table 1. Students with missing data on any of the six covariates (grade, gender, ethnicity, socioeconomic status defined by enrollment in the free/reduced lunch program, English language learners, and students with disabilities) were excluded from the dataset prior to the matching process. Frequencies of the covariates in the final sample (N=1726) are provided in Table 2 for both the treated and control groups prior to matching. Both groups had similar frequencies of covariates, indicating viable candidacy for the propensity score matching method.

Table 1. Summary of Covariate Labels, Descriptions, and Codes

Table 1. Summary of Covariate Labels, Descriptions, and Codes		
<u>Label</u>	<u>Description</u>	<u>Code</u>
Group	Participation in intervention	0 = Control 1 = Treated
Grade	Grade level	8 = 8 th grade 9 = 9 th grade
Gender	Gender	0 = Female 1 = Male
Ethnic	Ethnicity	1 = Asian 2 = Black 3 = Hawaiian/Pacific Islander 4 = Hispanic 5 = Non-disclosed 6 = White 7 = Native American
Lunch	Participates in Free/Reduced Lunch program; indicator of SES	0 = No 1 = Yes
ELL	Participates in services for English Language Learners	0 = No 1 = Yes
SWD	Student with Disabilities	0 = No 1 = Yes

Table 2. Frequencies of Covariates

Table 2. Frequencies of Covariates (N=1726)			
<u>Covariate Label</u>	<u>Covariate Option</u>	<u>Control</u>	<u>Treated</u>
Grade	8	472	445
	9	368	441
Gender	Female	403	452
	Male	437	434
Ethnic	Asian	40	28
	Black	333	335
	Hawaiian/Pacific Islander	0	1
	Hispanic	377	392
	White	83	126
Lunch	Native American	7	4
	No	66	112
ELL	Yes	774	774
	No	515	580
SWD	Yes	325	306
	No	729	802
	Yes	111	84

Similar to other program evaluation studies, the dependent variable (provided from the school district) for this dataset to assess impact of the intervention was cumulative grade point average (GPA). The variable cumulative GPA was scaled on a typical 4.0 scale with 0.0 being the lowest possible score and 4.0 being the highest. The table below displays the descriptive statistics for cumulative grade point average for the treated and the control group.

Table 3. Descriptive Statistics of GPA

Table 3. Descriptive Statistics of GPA				
<u>Group</u>	<u>N</u>	<u>Mean</u>	<u>Standard Deviation</u>	<u>Standard Error</u>
Control	840	2.609	0.850	0.029
Treated	886	2.716	0.854	0.0287

Software

The software packages used to complete this study include SPSS (IBM version 20.0) and R (R version 3.2.2 “Fire Safety”). The initial analysis of the dataset provided from the school district (e.g., frequencies, descriptive statistics, correlations, and t-tests) was calculated using SPSS. The R software program was used to assess the three research questions.

Few software packages are currently available for implementing propensity score matching. At this time, Stata (StataCorp, 2008), R (The R Foundation for Statistical Computing, 2008), and SAS offer computational packages for performing PSM. The R software package was used to compute the propensity score matching for this study based on its comprehensive and convenient offerings and the author’s experience with this software. R is a free statistical package that can be downloaded and installed from the website: <http://www.r-project.org>. There are various R packages that will conduct propensity score matching including MatchIt (Ho, King, & Stuart, 2011), Matching (Sekhon, 2011), twang (Ridgeway, McCaffrey, & Morral, 2006), cem (Iacus, King, & porro, 2008), optmatch (Hansen & Frederickson, 2009), PSAGraphics (Helmreich & Pruzek, 2009), and Synth (Abadie, Diamond, & Hainmueller, 2011). The most popular package in the R software program is MatchIt, based on this limited literature review, and was used in this research. The MatchIt: Nonparametric Preprocessing for Parametric Casual Inference software package in R was created by Ho, Imai, King & Stuart in 2011 in order to reduce model dependence in matching methods by preprocessing data with semi-parametric and non-parametric matching methods. This program allows researchers to use whatever parametric model and software they want without modifications.

Simulation Study

To explore the covariate selection problem with limited sample sizes, a Monte Carlo simulation study was created based on the real data sample. Unstandardized beta coefficients of reasonable size for educational research were selected based on the original data set. The treatment effect found in the initial analysis after the match was used to evaluate the precision of the estimates for the varying associations of the covariates to the outcome and to the treatment. One hundred replications were required for each condition in order to increase power of the findings. The averages of these replications were reported for each condition.

Research Questions

The considerations for conducting propensity score matching is ultimately a decision of which is superior: bias reduction or precision in matching. Because they are related, this decision places importance on one at the expense of the other. In matching, the removal of bias is preferred over the precision of the matches and is used to evaluate the quality and success of the method. Similarly, in this study the success of the propensity score matching method, as it relates to limitations in sample size, will be assessed by the removal of bias. However, both bias removal and precision in matching will be reported for each condition.

Many questions regarding best practices in propensity score matching for educational research remain. The primary question in educational research with policy evaluation is how well some intervention is meeting the intended goal. Quite often this question is convoluted in educational research as so many confounding variables can be present. Therefore, this question often becomes which method of analysis would be best in order to determine a causal effect using a quasi-experimental design. This research addresses how effective propensity score

matching is at reducing bias with the aforementioned common issues in educational research of sample size, caliper width, and covariate selection using a real life dataset compared to simulated data with the following questions.

Initial Analysis

Often in research with a quasi-experimental design, the true effect is unknown. In this research, the original dataset provided by the school district (N=1762) served as a baseline comparison or “true score” for all subsequent research questions. Some initial analyses were conducted on this original sample. First, simple mean differences, t –tests, were used to determine whether mean differences between the treated and control groups were statistically significant both before and after matching. This type of analysis is confounded with selection bias, but is frequently used in educational research and was therefore considered a benchmark for the results obtained in this study using propensity score matching.

Logistic regression was used to collapse the covariates into the propensity score, or the likelihood of having received the treatment. Using the following equation in which T_i is group membership (1=treatment), X_{1i}, \dots, X_{ni} are the scores on the covariates, and e_i is random error:

$$\ln\left(\frac{P_i(T_i = 1)}{1 - P_i(T_i = 1)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_n X_{ni} + e_i$$

The six covariates included in this analysis used to predict the propensity of a student to be included in the intervention were grade, gender, ethnicity, free/reduced lunch status, disability status, and English language learners. They were matched using the MatchIt program in the R software. Matches were created using the Nearest Neighbor method, which selects matches that

are nearest to the treated subject one at a time using a distance option with the order specified as largest. This type of matching is often called “greedy matching” because it matches treated subjects to controls without minimizing a distance measure. Participants were selected without replacement using one-to-one matching. Both of these techniques for matching, nearest neighbor and one-to-one matching, were chosen after reading a study by Austin who found that these techniques account for 83% of the methods used to conduct propensity score matching implemented by the studies between 1996 and 2003 (Austin, 2008).

As mentioned previously, a crucial part of any matching procedure is assessing the balance of the covariate distributions in the treatment group and the control group. The MatchIt software provided the standardized mean bias before and after the match for each covariate and summarizes this information into quantile-quantile plots (QQ plots), jitter plots, and histograms. The balance between the treated subjects and the control subjects was assessed both before and after matching. If the matching is successful, the measures of balance will be smaller in the matched data set. The results of the matching provides means, the original control group standard deviation, mean differences, standardized mean differences, and quantile-quantile plot differences in addition to the numerical overview of how many units were matched, unmatched and discarded and the percent improvement of balance after matching. The standardized mean difference was assessed because it is not confounded by sample size. A bias reduction in the standardized mean differences less than ten percent after matching was considered effective, based on this literature review of what is most frequently used. The c-statistic, used to assess the quality of the propensity scores and goodness of fit, is available using the rms package and was installed in addition to the MatchIt package. A c-statistic higher than 0.7 was considered favorable and greater than 0.8 strong based on this literature review (Hosmer et al., 2000).

Research Question 1: Limitations in Sample Size

How does propensity score matching perform with limitations in sample size as measured by bias reduction?

As mentioned previously, the need for a large sample size when conducting propensity score matching, to allow for overlap between the treated and control groups with many potential relevant covariates has been documented (Dehejia & Wahba, 2002; Fan & Nowell, 2011; Glazerman et al., 2003; Lane et al., 2012; Luellen et al., 2005; Rubin, 1997; Smith & Todd, 2005; Yanovitzky et al., 2005). It seems logical that in order to have sufficient matches, there must be a large pool of potential matches in the control group. Some research suggests that propensity score matching may only be feasible when there are extremely large sample sizes over 10,000 (Peikes et al., 2008). However, educational research is often fraught with low sample sizes, typically below 500. One study found that sample sizes less than 300 may be too small for matching when prediction of group assignment is high (Lane et al., 2012). The minimum requirement of sample size for propensity score matching is yet to be determined. The answer to this question can help researchers in education evaluate whether propensity score matching is a good method to use with limitations in sample size.

The full dataset, with a sample size of 1726 students (886 in the treated group and 840 in the control group), will be matched prior to sample reduction to pose as a “true score” of best possible bias reduction. Four sample sizes were chosen of 500, 400, 300, and 200 for comparison. The table below displays the four conditions for this study.

Table 4. Q1: Limitations in Sample Size

Table 4.
Q1: Limitations in Sample Size

<u>Conditions</u>	<u>N</u>
1	500
2	400
3	300
4	200

These four conditions were selected due to the lack of information found on sample sizes less than 500 regarding the efficacy of propensity score matching, coupled with the abundance of educational research constrained by limitations in sample size below this same level. More specifically, sample sizes above 300 were chosen to assess if previous findings of successful bias reduction using propensity score matching could be found using a real dataset in educational research. Sample sizes below 300 were chosen due to the studies on the lack of efficacy in bias reduction when using propensity score matching at these limited sample sizes.

Analysis

The four sample sizes were randomly selected from the dataset provided from the school district of 1726 students in the R software for comparison of bias reduction to the “true score” in the original dataset. This analysis was replicated 100 times for each condition of sample size in order to increase power and precision of the findings. The mean reduction in bias of these replications for each condition is presented.

Research Question 2: Caliper Width

What is the ideal caliper width used for propensity score matching when a sample is limited in size, as determined by bias reduction?

As mentioned previously, a caliper is the preset tolerance for the distance between the propensity scores of the treated and control subjects, which enables the matching (Rosenbaum & Rubin, 1984). The size of the caliper is determined a priori by the investigator, but at what tolerance the caliper width should be set is still in question. Most researchers use a caliper width 0.2 standard deviations of the logit of the propensity score due to research suggesting that this caliper size would reduce 99% of the bias in propensity scores (Austin, 2008, 2009; Austin et al., 2007) while others have used a caliper width of 0.6 as this same research stated this caliper width would reduce 90% of the bias (Ayanian et al., 2002). Propensity score matching has been used more frequently in medical research, where caliper widths can vary greatly.

With limitations in sample size, the question of what caliper width to use becomes even more complicated. A narrower caliper width will result in matching more similar subjects, reducing bias by reducing the systematic differences in the treated and control groups, but will also reduce the final matched sample size and will increase the variance in the estimated treatment effect. Using a wider caliper width will retain more subjects, decrease the variance in the estimated treatment effect, but will increase the bias and systematic differences in the two groups. Currently, there is not been an identified optimal caliper width as it relates to studies with limitations in sample size. The table below displays the four conditions of caliper width that were selected for this research, based on the literature review.

Table 5. Q2: Caliper Width

Table 5. Q2: Caliper Width	
<u>Condition</u>	<u>Caliper Width</u>
X ₁	0.1
X ₂	0.2
X ₃	0.3
X ₄	0.6

The propensity score matching studies that have been conducted in educational research thus far have consisted mostly of research using caliper widths of 0.2. However in medical research, where propensity score matching is far more utilized, calipers varied from .0005 to .6. The four conditions for caliper widths for this study were chosen based on frequency of use for both education and medical research.

Analysis

To address this question, the aforementioned four sample sizes that were randomly drawn from the original dataset (e.g., 500, 400, 300, 200) were matched with each of the four conditions of caliper widths (table above) for a total of 16 conditions displayed in Table 6 below.

Table 6. Q2: Caliper Width Conditions by Sample Size

Table 6. Q2: Caliper Width Conditions by Sample Size				
N	Caliper Width			
	0.1	0.2	0.3	0.6
500	X ₁	X ₂	X ₃	X ₄
400	X ₅	X ₆	X ₇	X ₈
300	X ₉	X ₁₀	X ₁₁	X ₁₂
200	X ₁₃	X ₁₄	X ₁₅	X ₁₆

This analysis was replicated 100 times for each condition of sample size in order to increase power and precision of the findings. The mean reduction in bias for these replications for each

condition is presented as well as the mean number of successful matches, exclusions, and c-statistics after the match.

Research Question 3: Covariate Selection

What relationship of the covariates to treatment and outcome is optimal when sample size is limited, as determined by bias reduction and treatment effect?

Research by Rubin found that including variables that are strongly related to the treatment but unrelated to outcome will decrease the efficiency of the estimated treatment effect; however, he also stated that excluding those variables would actually cause more concern with estimating the treatment effect than leaving them in the design when samples are substantial in size (1997). These results have been replicated by other studies with large sample sizes (Brookhart et al., 2006; Drake, 1993; Perkins et al., 2000; Robins et al., 1992). A Monte Carlo simulation conducted by Brookhart and his colleagues found that it would be advantageous to exclude covariates that are confounders in small studies ($n=500$) that are strongly related to the treatment and only weakly related to the outcome (2006). The inclusion of such a covariate and the increase in variance is not offset by a decrease in bias to improve the mean squared error. However, in this literature review, studies assessing the inclusion or exclusion of covariates as it relates to treatment and/or outcome with sample sizes less than 500 were not found.

To address this question, a Monte Carlo simulation was created in the R software program based on the real dataset. The unstandardized beta coefficients (b) of the covariates were observed in the original dataset as a baseline of typical coefficients found in educational research. A quick literature review was also conducted to assess typical relationships of covariates to treatment and outcome in the education field. Unstandardized beta coefficients

were used to avoid possible bias due to standardization with small samples. This coefficient (b) indicates the average change in the dependent variable (e.g., final GPA) associated with one unit change in the corresponding predictor covariate, while statistically controlling for the other independent variables (covariates). The unstandardized beta coefficients chosen for each relationship to outcome were defined as no relationship = 0, weak relationship = .20, moderate relationship = .40, and strong relationship = .60. Odds ratios were used to determine the measure of association between the covariates and the treatment. The odds ratio represents the odds that an outcome will occur, given a particular exposure, compared to the odds of the outcome in the absence of that exposure. In other words, the odds ratio will be the odds that a student would be in the treatment group, given the covariates in this study (grade, gender, ethnicity, free/reduced lunch status, ELL status, and SWD status). Odds ratios greater to one are interpreted as exposure of the covariate is associated with higher odds of the outcome (being in the treated group); ratios less than one mean that exposure is associated with lower odds of the outcome. Odds ratios equivalent to one are interpreted as exposure does not affect the odds of the outcome. The magnitude of the odds ratios chosen for association of the covariates to the treatment assignment were defined as no relationship = 1, weak relationship = 1.44, moderate = 2.47, strong = 4.25 as determined from Chinn (2000). The simulation was designed to assess what combinations of association of covariates to the treatment were optimal when the relationship of the covariates to outcome were strong and vice versa. The average treatment effect was set at 0.11, based on the findings of the original dataset. Conditions for the associations of the covariates to outcome and treatment were chosen at the extremes. In other words one relationship was held constant as a strong association while the other associations were evaluated. The table below shows the six conditions that were assessed for covariate relationship to outcome and to treatment.

Table 7. Q3. Covariate Selection Relations

Table 7. Q3. Covariate Selection Relations		
<u>Condition</u>	<u>Relation to Outcome</u>	<u>Relation to Treatment</u>
X_1	Strong	None
X_2	Strong	Weak
X_3	Strong	Moderate
X_4	None	Strong
X_5	Weak	Strong
X_6	Moderate	Strong

Sample sizes (500,400,300,200) and caliper widths (0.1, 0.2, 0.3, 0.6) designated in the previous questions of this study will be utilized for this question as well. The same criterion for the match, one-to-one matching using the nearest neighbor method without replacement, was used. This analysis was replicated 100 times for each condition in order to increase power and precision of the findings. The mean reduction in bias for these replications for each condition is presented as well as the mean number of successful matches, exclusions, and c-statistics after the match. The impact on measurement of the “true” treatment effect given the limitations in sample size is also discussed.

Data Analysis Summary

Overall, limitations in sample size were assessed as it related to covariate selection and caliper width. The original dataset served as a baseline or “true score” for best possible bias reduction. Sample sizes were randomly reduced from this original dataset to assess bias reduction as it related to caliper width. These analyses were replicated 100 times and the mean statistics are provided for each condition of sample size and caliper width. A Monte Carlo simulation was created based on the original dataset to assess covariate selection as it related to

the strength of the relationship to the treatment and to the outcome with the same limitations in sample size and conditions for caliper width. These analyses were also replicated 100 times in the R software program to increase power and precision of the findings. A complete copy of the R code used to complete this analysis is available in Appendix C.

Chapter Four – Results

The propensity score matching technique has considerable potential for educational research. The current research on propensity score matching is missing imperative information for educational researchers regarding the practical implications of utilizing this method with sample sizes smaller than 500. The objective of the current study was to assess the effectiveness of propensity score matching at reducing bias in a real data set with limited sample sizes and varying caliper widths as compared to an ideal data set created in a Monte Carlo simulation study. It was hypothesized that the effectiveness of propensity score matching at reducing bias will drop below acceptable levels for research when sample size is reduced to a certain point.

Initial Analysis

The initial dataset from the school district served as a baseline comparison for the research questions in this study. Participants with missing data in any of the covariates or the dependent variable of cumulative GPA were removed. The remaining dataset included 1726 students to be matched (886 in the treatment group and 840 in the control group). The six covariates included in this analysis used to predict the propensity of a student to be included in the intervention were grade, gender, ethnicity, free/reduced lunch status, disability status, and English language learners. The dependent variable provided for this dataset in order to evaluate impact of the intervention was cumulative GPA. Pearson correlations of these covariates to each other, to the treatment assignment, and to the outcome of cumulative GPA were conducted in order to evaluate the contribution of each covariate and can be found in the table below.

Table 8. Correlations of Covariates to Treatment and Outcome

	Grade	Gender	Ethnicity	Lunch	ELL	SWD	Treatment	Final GPA
Grade	1.0							
Gender	.018	1.0						
Ethnicity	-.011	.009	1.0					
Lunch	-.086**	-.016	.137**	1.0				
ELL	.013	.057*	.263**	.119**	1.0			
SWD	-.023	.072**	-.064**	-.017	-.032	1.0		
Treatment	.060*	-.030	.056*	-.079**	-.043	-.059*	1.0	
Final GPA	.005	.007	-.034	-.023	-.005	.044	.063**	1.0

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

In the table, the variable ‘Treatment’ identifies the treatment assignment and ‘Final GPA’ is the outcome variable. Pearson’s correlation coefficient ranges from -1 to +1. Interpretation of the strength of the correlations used Cohen’s 1988 descriptions (weak $0.1 < |r| < 0.3$; moderate $0.3 < |r| < 0.5$; strong $0.5 < |r| < \dots$). Using these benchmarks, the strengths of the correlations of the covariates to each other, to the treatment assignment, and to the outcome are considered weak. Of specific interest is the relationship of the covariates to the treatment assignment when choosing which covariates to include in a matching study. The covariates with a significant relationship to the treatment assignment were ethnicity (.056, $p < .05$), students with disabilities (-.059, $p < .05$), grade level (.060, $p < .05$), GPA (.063, $p < .01$), and free/reduced lunch status (-.079, $p < .01$). These findings are not surprising, given that the requirement to be included in the intervention was at-risk students in a specific grade level. The determination of treatment assignment was significantly correlated with the outcome variable of Final GPA (.063, $p < .01$). This significance identifies a positive relationship between students being involved in the intervention and the outcome variable of final grade point average.

The covariates included in this study are all nominal variables; they are categorical. The outcome, dependent, variable of final grade point average is considered to be a ratio variable because it has a true zero on the scale of 0 to 4.0. A correlation indicates the strength and direction of a linear relationship among continuous, not categorical variables. Therefore, a chi-square test was also conducted for the covariates to assess the relationship of the covariates to the treatment assignment. The chi-square statistic evaluates the impact of a covariate by comparing how the log-likelihood decreases or increases as covariates are added or deleted. The table below provides the results of the chi-square tests.

Table 9. Chi-Square Test of Covariates

Table 9. Chi-Square Test of Covariates (N=1726)						
Covariate	Covariate Option	Control	Treated	Chi-Square	p	Cramer's V
Grade	8	472	445	6.16	.01*	.06
	9	368	441			
Gender	Female	403	452	1.59	.21	.03
	Male	437	434			
Ethnic	Asian	40	28	11.86	.04*	.08
	Black	333	335			
	Hawaiian/Pacific Islander	0	1			
	Hispanic	377	392			
	White	83	126			
	Native American	7	4			
Lunch	No	66	112	10.67	p<.001**	.08
	Yes	774	774			
ELL	No	515	580	3.21	.07	.04
	Yes	325	306			
SWD	No	729	802	6.00	.01*	.06
	Yes	111	84			

*Significant at the 0.05 alpha level

**Significant at the 0.01 alpha level

An alpha level of .05 was used to determine a significant association between categorical variables for the chi-square tests. The chi-square test determined whether a relationship between

the covariate and treatment assignment exists. Greater differences between an expected and actual data produce a larger chi-square value. The larger the chi-square value, the greater the probability that there is a significant difference. Similar to the findings of the Pearson correlations, significant differences were found for the covariates grade (6.16, $p < .05$), ethnicity (11.86, $p < .05$), free/reduced lunch status (10.67, $p < .01$), and students with disabilities (6.00, $p < .05$).

If the relationship between a covariate and the treatment assignment was found to be significant, the Cramer’s V indicated the strength of that significance. Cramer’s V values can range from 0 to 1. Classifications for Cramer’s V statistics utilized for this study were taken from Cohen’s suggestion (small = 0.1, medium = 0.3, large = 0.5). Using these benchmarks, all covariates with significant chi-square statistics were found to have small Cramer’s V values.

An independent samples t –test was calculated to determine whether mean differences on the outcome variable of grade point average between the treated and control groups were statistically significant before matching. Simple mean differences between groups alone, both before and after an intervention, are not a sufficient indicator of impact of an intervention. However, this method of analyzing impact is typical and was included in the analysis as one indicator of potential matching success. The table below displays the findings from the independent samples t-test before matching.

Table 10. Independent Samples t-test for GPA before match

Table 10. Independent Samples t-test for GPA								
<u>Group</u>	<u>N</u>	<u>Mean</u>	<u>Standard Deviation</u>	<u>t</u>	<u>df</u>	<u>p</u>	<u>Mean Diff</u>	<u>Standard Error</u>
Control	840	2.61	0.85	-2.607	1724	.009*	-.107	.041
Treated	886	2.72	0.85					

*Significant at the 0.05 alpha level

Findings from the independent samples t-test indicate a significant difference on the outcome variable final grade point average between the treated and control groups. More specifically, the mean grade point average of the control group is significantly less than the treated group by .107 points. This significant difference between groups provides evidence in support of utilizing the propensity score matching method.

Initial Match

The participants were matched using the MatchIt program in the R software. Matches were created using the most common methods, the Nearest Neighbor method, without replacement, using one-to-one matching with a caliper width of 0.2. The dataset included 1726 students prior to the match, with 886 in the treated group and 840 in the control group. After the students were matched the sample included 1478 students, with 739 in both groups. Frequencies of the samples before and after they were matched are provided in the table below.

Table 11. Frequency of the Covariates before and after Match

Covariate		Before Match		After Match	
		N=1726		N=1478	
		Control	Treated	Control	Treated
Grade	8	472	445	398	404
	9	368	441	341	335
Gender	Female	403	452	369	375
	Male	437	434	370	364
Ethnic	Asian	40	28	26	24
	Black	333	335	291	295
	Hawaiian/Pacific Islander	0	1	0	1
	Hispanic	377	392	334	329
	White	83	126	82	90
	Native American	7	4	6	0
Lunch	No	66	112	58	59
	Yes	774	774	681	680
ELL	No	515	580	469	478
	Yes	325	306	270	261
SWD	No	729	802	665	662
	Yes	111	84	74	77

Matching without replacement, with the ratio of one-to-one and a caliper width of 0.2, 85.6% of the original sample was successfully matched and 14.4% had to be excluded (101 control and 147 treated). The findings used to evaluate the success of the matching method, provided in the MatchIt software, are provided in the table below.

Table 12. Summary of Balance for Matched Data

Table 12. Summary of Balance for Matched Data (N=1478)					
<u>Covariate</u>	<u>Before Match</u> <u>Standardized Mean</u> <u>Difference</u>	<u>After Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>Balance</u> <u>Improvement</u>	<u>Successfully</u> <u>Matched</u>	<u>Excluded</u>
Overall	.02	.00	99.98%	85.6% (N=1478)	14.4% (N=248)
Grade	.06	-.00	86.39%		
Gender	-.03	-.00	73.29%		
Ethnic	.16	-.00	95.74%		
Lunch	-.05	-.00	97.17%		
ELL	-.04	.00	70.68%		
SWD	-.04	-.01	89.13%		

The matching process for this initial dataset of 1726 participants resulted in many matches and a substantial reduction in bias overall. The c-statistic was observed as a discrimination index of the logistic regression for how well the model can discriminate between observations at different levels of the outcome. The minimum value of a c-statistic is 0.5 and the maximum is 1.0. In their textbook, Hosmer and Lemeshow consider c-statistic values of 0.7 to 0.8 to show acceptable discrimination, values of 0.8 to 0.9 to indicate excellent discrimination, and values of ≥ 0.9 to show outstanding discrimination. The c-statistic value in the unadjusted model was 0.576 and 0.512 in the adjusted model, both below the threshold for acceptable discrimination, yet the decrease in c-statistic indicates better prediction of the model after the match due to possible exclusion of outliers. These findings will serve as a “true score” of best possible indices as sample size is reduced and caliper width is varied in this study.

To further evaluate the quality of the original match, the MatchIt software provides graphical representations of the distributions of the propensity scores before and after the match (e.g., jitter plots and histograms). The figure below is the jitter plot of the original match.

Distribution of Propensity Scores

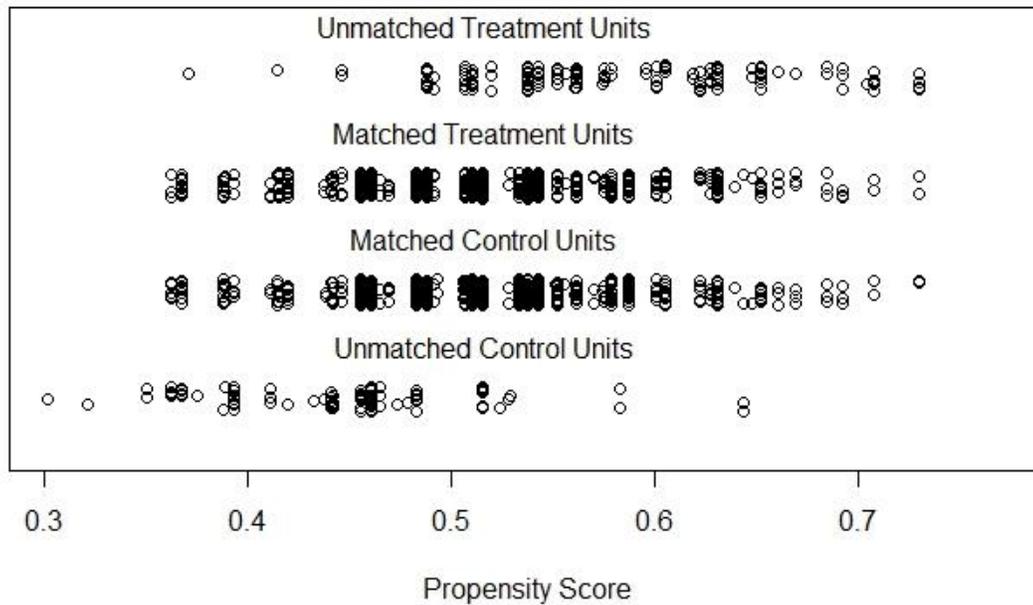


Figure 7. Jitter Plot of Original Data (N=1762)

Jitter plots display the overall distribution of propensity scores in the treated and control groups. The size of each point is proportional to the weight given to that unit. The desired outcome of successful matching is to visually observe similar distributions of the propensity scores for the matched treated and control groups. In the figure above, you can see the distributions for the propensity scores prior to match for the treated subjects is heavily distributed to the right and the treated subjects are heavily distributed to the left. After the match, both groups have similar distributions of propensity scores, suggesting successful matches of the two groups. Another visual representation of the distribution of the propensity scores before and after the match is a histogram, provided below.

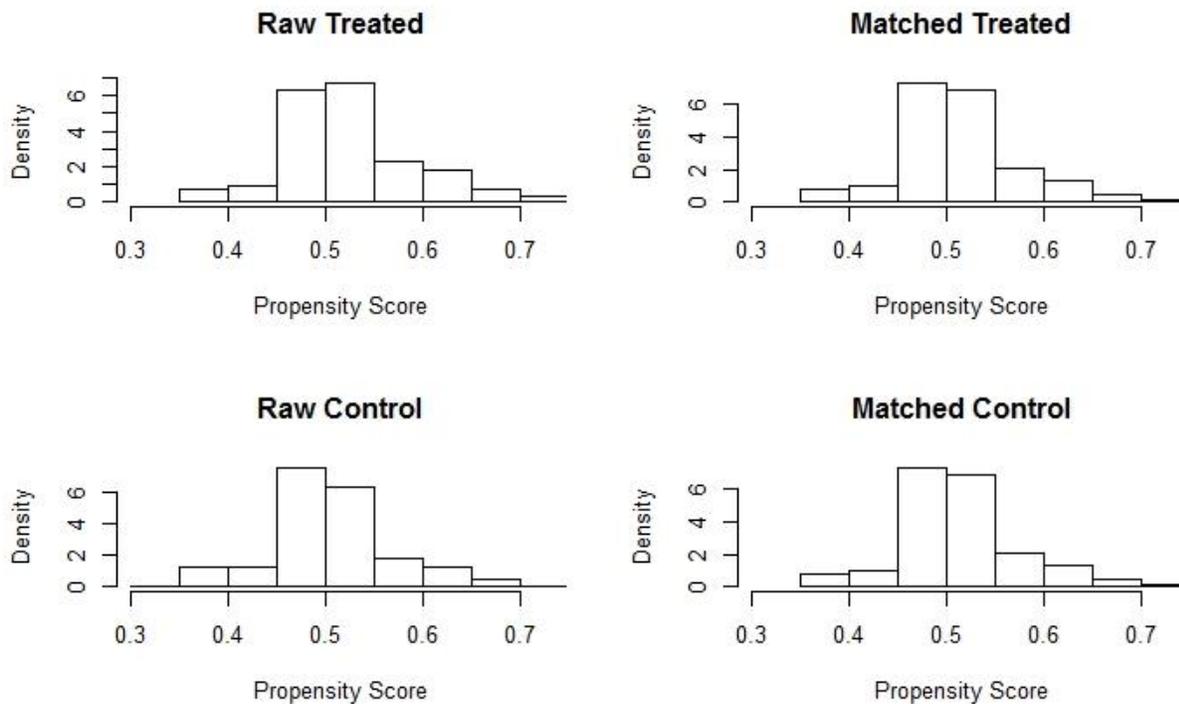


Figure 8. Histogram of the Original Data (N=1762)

These histograms can be compared vertically in order to quickly assess the balance before and after matching have been completed. After matching, the histogram distributions of the propensity scores on the right are more similar than prior to matching, indicating successful matching.

The following figures 9 through 12 display the distributions of the outcome variable, Final GPA, before and after the match of the original dataset.

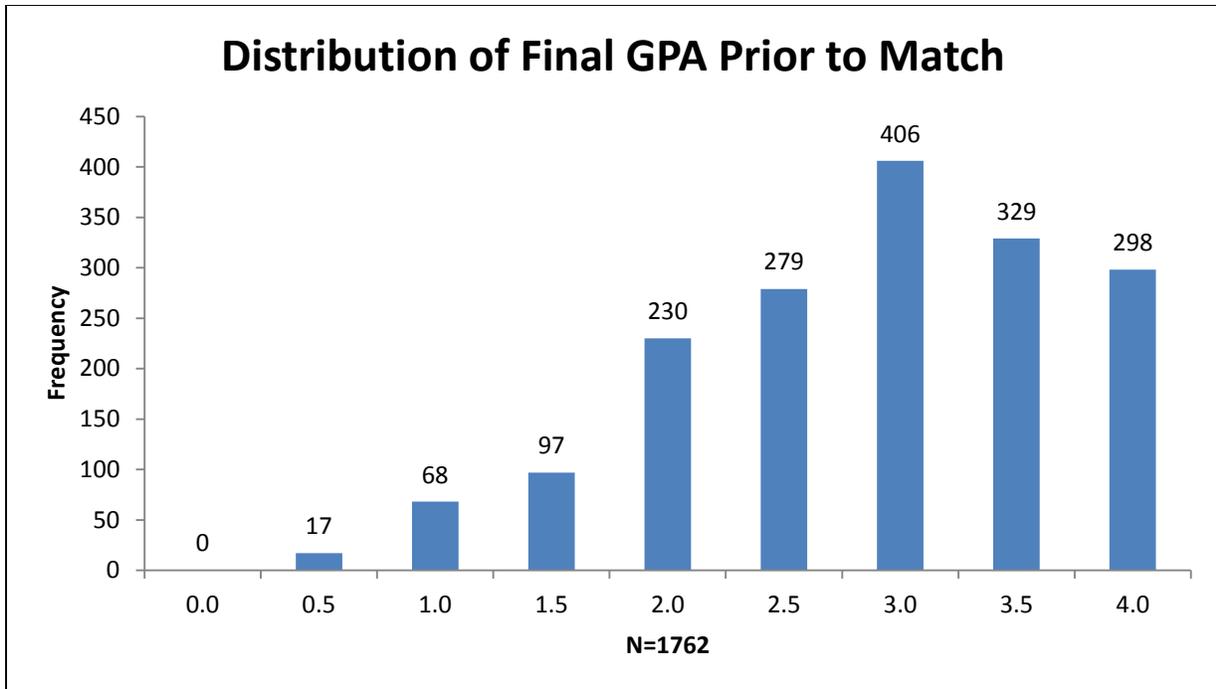


Figure 9. Distribution of Final GPA Prior to Match

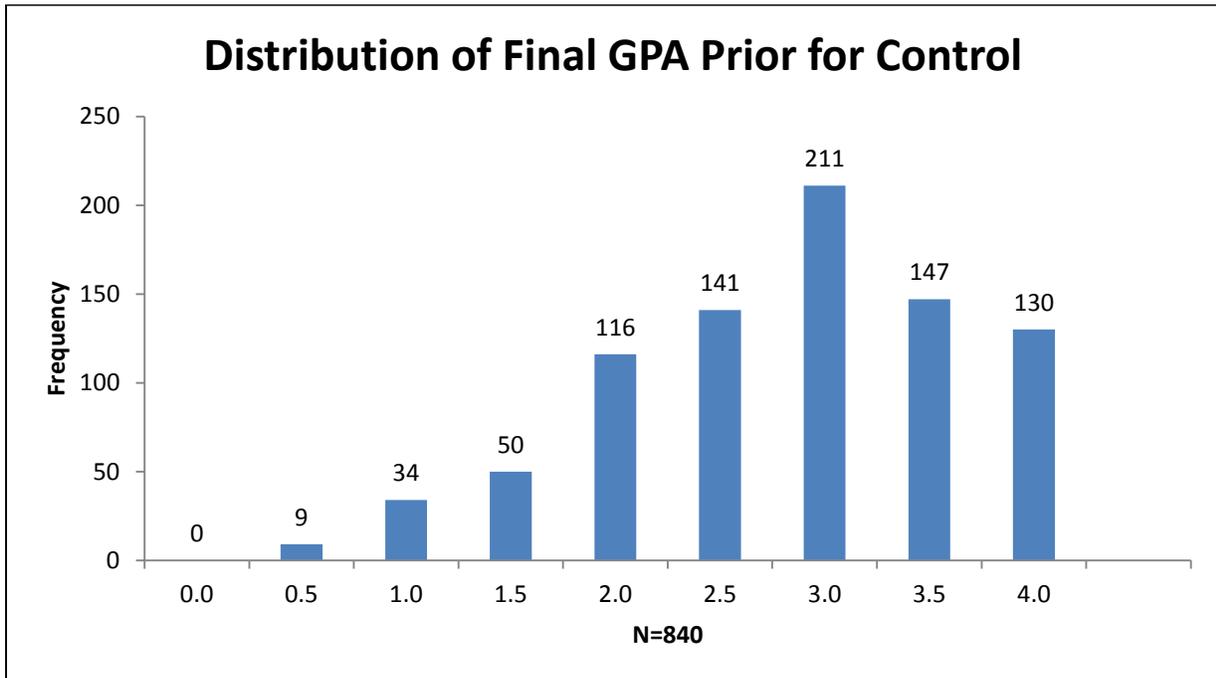


Figure 10. Distribution of Final GPA Prior for Control

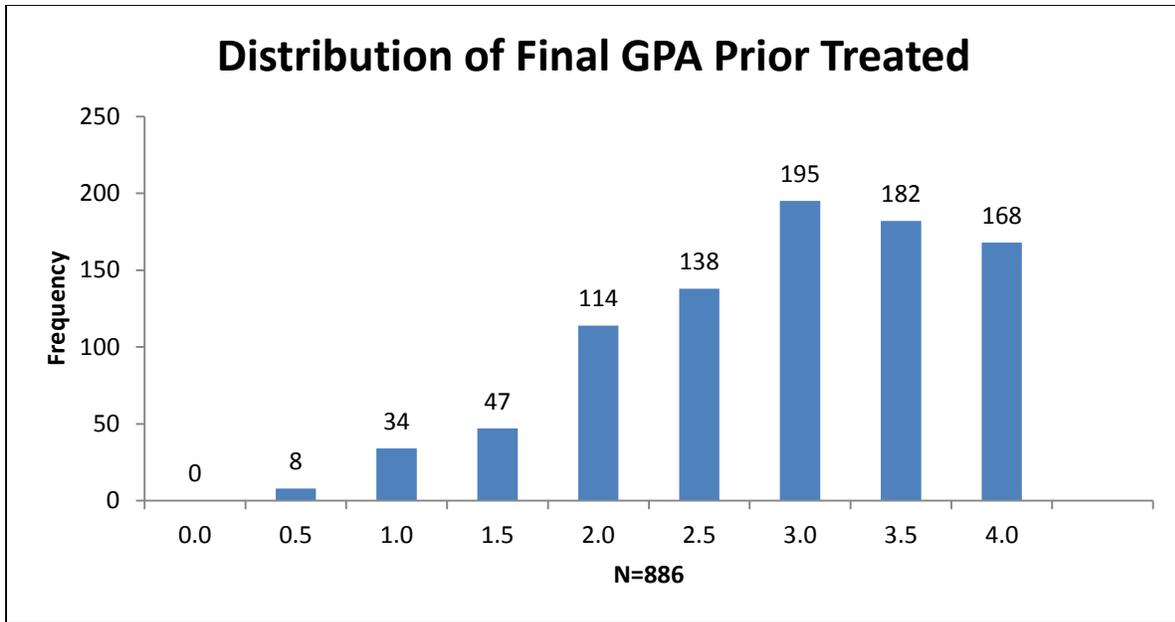


Figure 11. Distribution of Final GPA Prior Treated

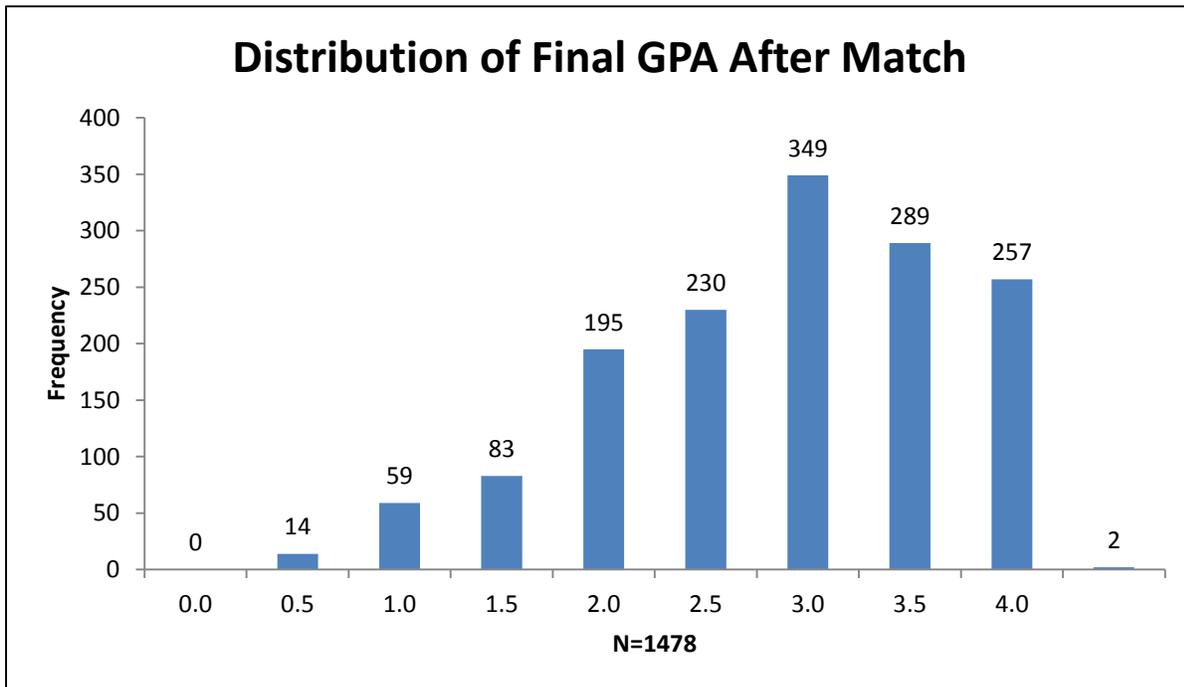


Figure 12. Distribution of Final GPA after Match

These figures display the distributions of the outcome variable before and after the match. All distributions are negatively skewed, with more students above a 3.0 GPA in the treated group (61.5%) compared to the control group (58.1%) prior to the match. After the match, the distribution of the outcome variable, Final GPA, is negatively skewed but more even distributed.

After the match was completed for the initial analysis, an independent samples t-test was conducted again in order to estimate the possible treatment effect on cumulative grade point average. The table below displays the findings from the independent samples t-test after matching.

Table 13. Independent Samples t-test for GPA after match

Table 13. Independent Samples t-test for GPA								
<u>Group</u>	<u>N</u>	<u>Mean</u>	<u>Standard Deviation</u>	<u>t</u>	<u>df</u>	<u>p</u>	<u>Mean Diff</u>	<u>Standard Error</u>
Control	739	2.59	0.87	-2.174	1476	.03*	-.099	.046
Treated	739	2.69	0.89					

*Significant at the 0.05 alpha level

Findings from the independent samples t-test after the match indicates a significant difference on the outcome variable, final grade point average, between the matched treated and control groups. The mean grade point average for the control group is significantly less than the treated by .11 points at the 0.05 alpha level. These findings will serve as a baseline for the following research questions:

Research Questions 1 and 2: Limitations in Sample Size and Caliper Width

How does propensity score matching perform with limitations in sample size as measured by bias reduction? What is the ideal caliper width used for propensity score matching when a sample is limited in size, as determined by bias reduction?

Research questions 1 and 2 were completed using this same dataset provided by the school district. Four sample sizes (500, 400, 300, and 200) were randomly drawn from the original dataset and were matched with each of the four conditions of caliper widths (0.1, 0.2, 0.3, and 0.6) for a total of 16 conditions. These conditions were replicated 100 times to increase the power and precision in the findings. The tables below display the means over all of the replications for the standardized differences in the means before and after the match, the balance improved (bias reduction), and the number of successful matches and exclusions, given the criterion selected.

Table 14. Summary of Balance for Matched Data (N=500)

Table 14. Summary of Balance for Matched Data (N=500)						
<u>Caliper Width</u>	<u>Before Match Standardized Mean Difference</u>	<u>After Match Standardized Mean Difference</u>	<u>Balance Improvement</u>	<u>Successfully Matched</u>	<u>Excluded</u>	
0.1	0.3138	0.0002	99.93%	38.8% (N=194)	61.2% (N=306)	
0.2	0.3303	0.0008	99.77%	39.0% (N=195)	61.0% (N=305)	
0.3	0.3279	0.0018	99.45%	39.6% (N=198)	60.4% (N=302)	
0.6	0.3047	0.0058	98.17%	40.8% (N=204)	59.2% (N=296)	

For the sample size of 500 you can see that regardless of the caliper width chosen, the balance improvement (or bias reduction) is all well above the acceptable level of 90%. The level

of balance improvement decreases as caliper width is widened, yet all are acceptable. Levels of balance improvement compare to that found in the original dataset matched (N=1478) with a balance improvement of 99.98%. For this dataset, this suggests that with samples of this size the choice of caliper width is insignificant because all have acceptable bias reduction. As expected as the caliper width becomes wider, the number of participants successfully matched increases, while the number of exclusions decreases. However, these numbers only slightly change when the caliper width is varied.

Table 15. Summary of Balance for Matched Data (N=400)

Table 15. Summary of Balance for Matched Data (N=400)					
<u>Caliper Width</u>	<u>Before Match Standardized Mean Difference</u>	<u>After Match Standardized Mean Difference</u>	<u>Balance Improvement</u>	<u>Successfully Matched</u>	<u>Excluded</u>
0.1	0.3325	0.0002	99.92%	37.8% (N=151)	62.3% (N=249)
0.2	0.3221	0.0008	99.77%	38.5% (N=154)	61.5% (N=246)
0.3	0.3452	0.0025	99.33%	38.8% (N=155)	61.3% (N=245)
0.6	0.3480	0.0092	97.53%	40.3% (N=161)	59.8% (N=239)

The sample size of 400 has similar trends, with all possible caliper widths creating balance improvement levels that are deemed acceptable (>90%). These levels of balance are also all comparable to that found in the original dataset matched (N=1478) with a balance improvement of 99.98%. The same expected trend for caliper width related to number of participants successfully matched versus excluded is seen.

Table 16. Summary of Balance for Matched Data (N=300)

Table 16. Summary of Balance for Matched Data (N=300)					
<u>Caliper Width</u>	<u>Before Match Standardized Mean Difference</u>	<u>After Match Standardized Mean Difference</u>	<u>Balance Improvement</u>	<u>Successfully Matched</u>	<u>Excluded</u>
0.1	0.3594	0.0003	99.92%	35.7% (N=107)	64.3% (N=193)
0.2	0.3315	0.0008	99.75%	36.7% (N=110)	63.3% (N=190)
0.3	0.3814	0.0028	99.29%	37.3% (N=112)	62.7% (N=188)
0.6	0.3640	0.0078	97.99%	39.7% (N=119)	60.3% (N=181)

This sample size of 300 also displays all possible caliper widths creating balance improvement levels that are deemed acceptable (>90%). The level of balance improvement decreases as caliper width is widened, yet all are acceptable. For this dataset, this suggests that with samples of this size the choice of caliper width is insignificant because all have acceptable bias reduction. These levels of balance are also all comparable to that found in the original dataset matched (N=1478) with a balance improvement of 99.98%. The same expected trend for caliper width related to number of participants successfully matched versus excluded is seen.

Table 17. Summary of Balance for Matched Data (N=200)

Table 17. Summary of Balance for Matched Data (N=200)					
<u>Caliper Width</u>	<u>Before Match Standardized Mean Difference</u>	<u>After Match Standardized Mean Difference</u>	<u>Balance Improvement</u>	<u>Successfully Matched</u>	<u>Excluded</u>
0.1	0.4054	0.0004	99.88%	33.5% (N=67)	66.5% (N=133)
0.2	0.4018	0.0013	99.69%	35.0% (N=70)	65.0% (N=130)
0.3	0.4108	0.0037	99.13%	35.5% (N=71)	64.5% (N=129)
0.6	0.4128	0.0146	96.77%	37.5% (N=75)	62.5% (N=125)

Unexpectedly, similar trends are again seen for this sample size of 200. All possible caliper widths for this sample size condition created balance improvement levels that are deemed acceptable (>90%). The levels of balance are all still comparable to that found in the original dataset matched (N=1478) with a balance improvement of 99.98%. The same expected trend for caliper width related to number of participants successfully matched versus excluded is seen for the sample size of 200. The figure below displays the overall bias reduction for all conditions of sample size and caliper width.

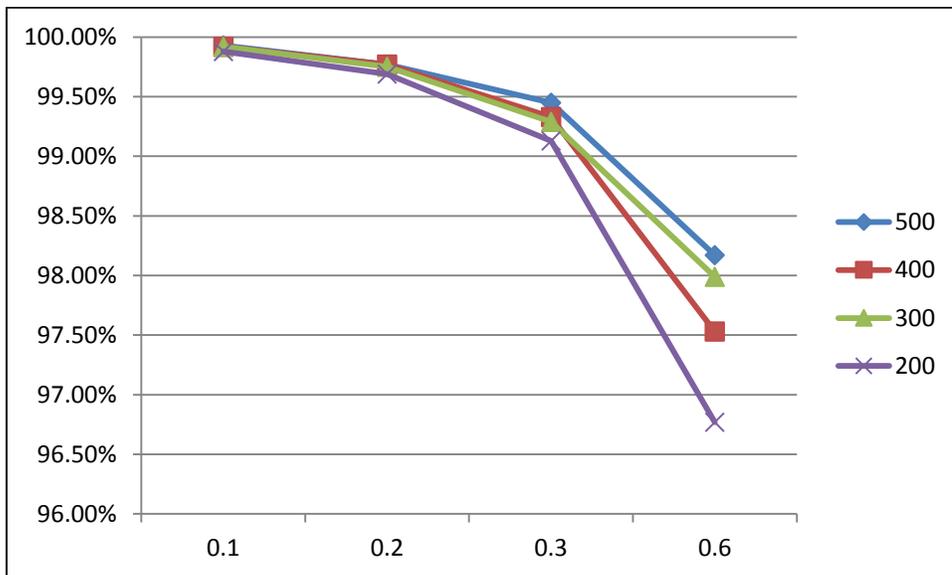


Figure 13. Bias Reduction by Sample Size and Caliper Width

As you can see in the figure, reduction in bias performs similarly for each condition of reduction in bias and caliper width until you get to the widest caliper of 0.6. At this point, the bias reduction varies from 98.17% for a sample size of 500 to 96.77% for a sample size of 200. Although each caliper width provided balance improvement at acceptable levels, the widest caliper one could chose before much dispersion would be 0.3.

Although the success of a matching method is not evaluated by the number of successful matches or exclusions, this information becomes more pertinent as samples are limited in order

to have a powerful comparison. The figures below display the number of successful matches and exclusions over all sample sizes and caliper widths. These reveal the trend discussed previously of greater retention of participants with wider calipers. As expected the greater the sample size, the greater the number of successful matches and fewer exclusions.

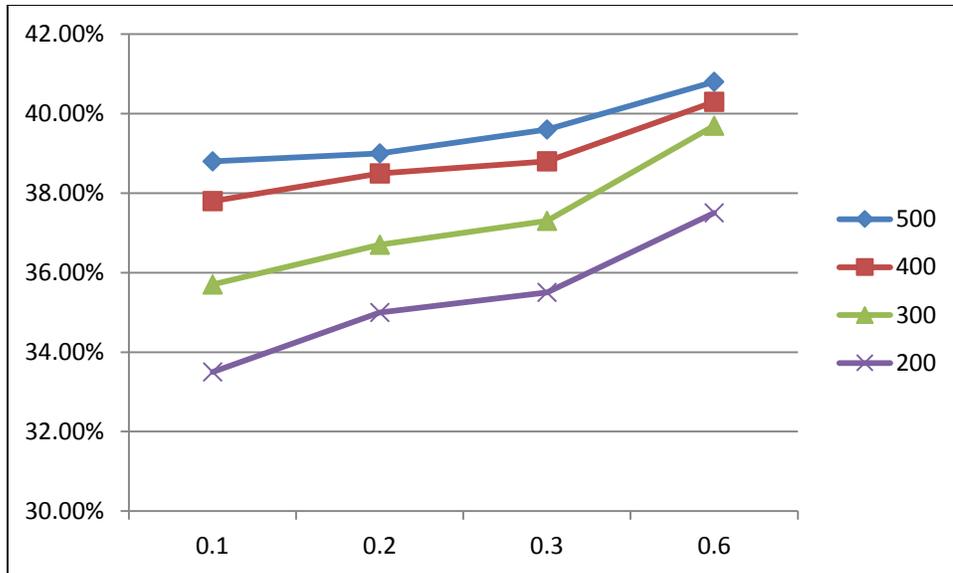


Figure 14. Match Success by Sample Size and Caliper Width

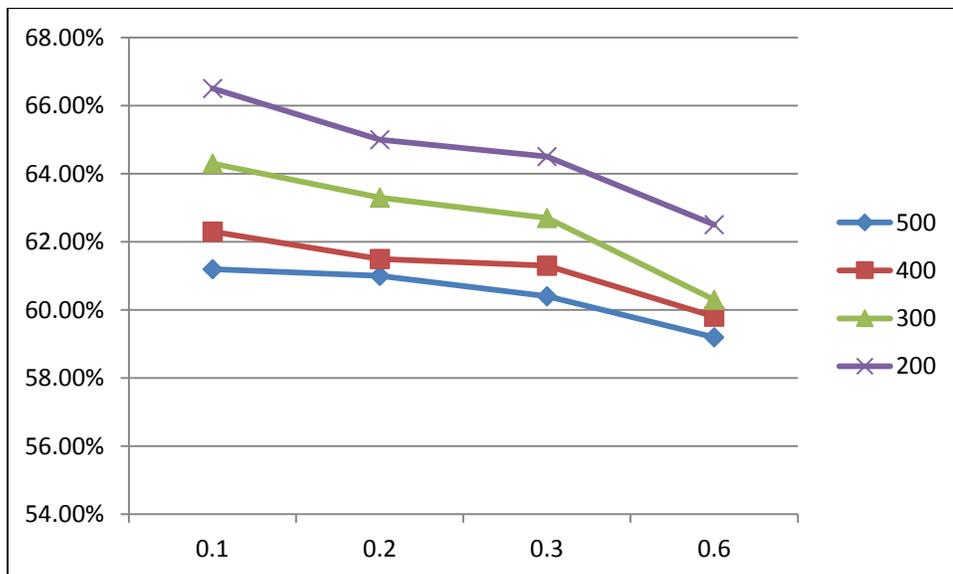


Figure 15. Exclusions by Sample Size and Caliper Width

The c-statistic is frequently used to assess the quality of the propensity scores and goodness of fit, in addition to the standardized mean bias reduction. It is a measure of the predictive ability of the model by calculating the proportion of the pairs in which the treated subject had a higher estimated propensity score than the control subject (Harrell, 2001). Values for this measure range from 0.5 to 1.0. A value of 0.5 indicates that the model is no better than chance at making a prediction of membership in a group, and a value of 1.0 indicates that the model perfectly identifies those within a group and those not. Models are typically considered reasonable when the c-statistic is higher than 0.7 and strong when c exceeds 0.8 (Hosmer et al., 2000). The mean c-statistic after the match for all of the 16 conditions is provided in the table below.

Table 18. Mean c-statistic by Sample Size and Caliper Width

Table 18. Mean c-statistic by Sample Size and Caliper Width				
<u>Caliper Width</u>	<u>Sample Size</u>			
	500	400	300	200
0.1	0.5211	0.5238	0.5258	0.5331
0.2	0.5270	0.5278	0.5303	0.5485
0.3	0.5388	0.5380	0.5443	0.5526
0.6	0.5450	0.5487	0.5528	0.5680

The c-statistic for the original dataset was considered below the threshold for what is considered to be reasonable or strong prior to matching ($c = .576$). After randomly reducing the sample size for the four conditions and matching across the four caliper conditions, the mean c-statistic ranged from 0.52 to 0.57. These values remain below the threshold of what is considered reasonable or strong for predictive power of a model by Hosmer and colleagues (2000). However, it is important to note that using a wider caliper width consistently increased the c-statistic for this study. These findings are consistent with previous studies that state that wider

caliper widths increase the power of the estimated treatment effect by decreasing the variance.

As calipers are widened the subjects matched are less similar and the higher the proportion of the pairs in which the treated subject had a higher estimated propensity score than the control subject. The figure below further reveals this trend.

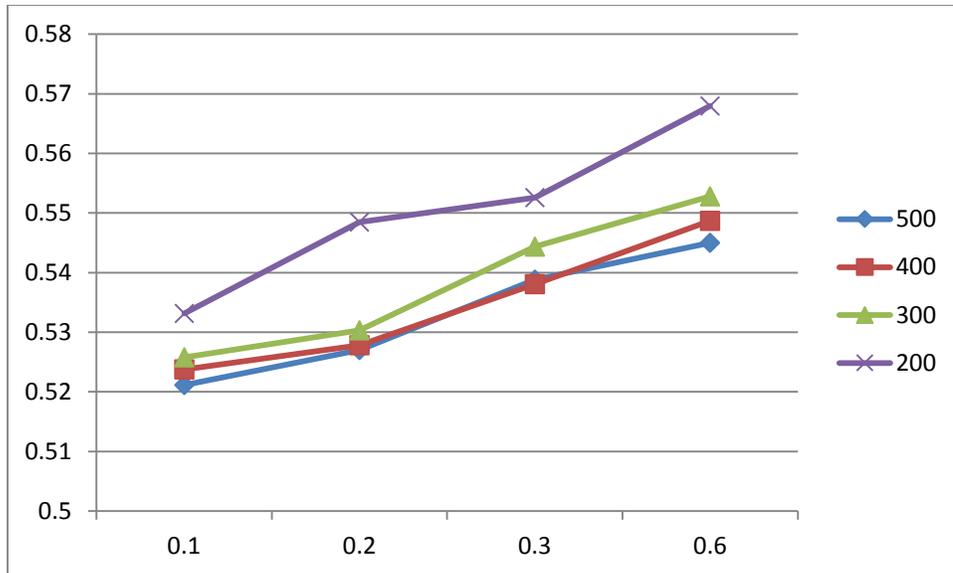


Figure 16. Mean c-statistic by Sample Size and Caliper Widths

Research Question 3: Covariate Selection

What relationship of the covariates to treatment and outcome is optimal when sample size is limited, as determined by bias reduction?

Research question 3 was completed using a Monte Carlo simulation was created in the R software package mimicking the original dataset used in the previous research questions. The same four sample sizes (500,400,300,200) were used to assess the associations of the covariates to the treatment and to the outcome of final grade point average. The strengths of these

associations were designated based on the associations in the real dataset in the previous questions. The unstandardized beta coefficients chosen for each relationship to outcome were defined as no relationship = 0, weak relationship = .20, moderate relationship = .40, and strong relationship = .60. The magnitude of the odds ratios chosen for association of the covariates to the treatment assignment were defined as no relationship = 1, weak relationship = 1.44, moderate = 2.47, strong = 4.25. Based on the findings in the previous study, it was decided to use a caliper width of 0.3 given that it maintained similar bias reduction regardless of sample size. However, for all sample sizes the caliper width of 0.3 was too stringent and the simulation would not converge. The convergence rates for each sample size were: 500 = 49%, 400 = 67%, 300 = 52%, and 200 = 6%. The caliper width of 0.6 was attempted and displayed no convergence issues with these limited sample sizes. Therefore, a caliper width of 0.6 was used to answer this research question. Findings from the previous questions indicated that this caliper width still had acceptable levels of bias reduction regardless of sample size. The “true” treatment effect of 0.11, based on the original dataset, was used in this simulation to evaluate how well the treatment effect was estimated given the varied covariate relationships and sample size. These conditions were replicated 100 times to increase the power and precision in the findings. Tables 19 through 22 below display the means over all of the replications for the standardized differences in the means before and after the match, the balance improved, and the number of successful matches and exclusions, given the criterion selected. The covariate associations are stated as the magnitude of the relationship to the outcome and then the relationship to treatment second (i.e., strong_none means a strong covariate/DV relationship and not covariate/treatment relationship).

Table 19. Summary of Balance for Simulated Match (N=500)

Table 19. Summary of Balance for Simulated Match (N=500)					
	<u>Before Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>After Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>Balance</u> <u>Improvement</u>	<u>Successfully</u> <u>Matched</u>	<u>Excluded</u>
strong_none	0.0672	0.0048	94.38%*	28.8% (N=144)	71.2% (N=356)
strong_weak	0.3625	0.0478	87.20%	41.0% (N=205)	59.0% (N=295)
strong_mod	0.8457	0.1693	80.22%	36.4% (N=182)	63.6% (N=318)
none_strong	1.2596	0.3336	73.68%	31.6% (N=158)	68.4% (N=342)
weak_strong	1.2645	0.3361	73.61%	31.8% (N=159)	68.2% (N=341)
mod_strong	1.2591	0.3360	73.50%	31.6% (N=158)	68.4% (N=342)

*Defined as effective bias reduction in the literature
All replications were completed with a caliper width = 0.6

Table 20. Summary of Balance for Simulated Match (N=400)

Table 20. Summary of Balance for Simulated Match (N=400)					
	<u>Before Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>After Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>Balance</u> <u>Improvement</u>	<u>Successfully</u> <u>Matched</u>	<u>Excluded</u>
strong_none	0.0721	0.0049	94.41%*	27.3% (N=109)	72.8% (N=291)
strong_weak	0.3734	0.0454	88.28%	40.0% (N=160)	60.0% (N=240)
strong_mod	0.8369	0.1609	81.04%	36.3% (N=145)	63.8% (N=255)
none_strong	1.2676	0.3293	74.22%	31.3% (N=125)	68.8% (N=275)
weak_strong	1.2682	0.3309	74.09%	31.3% (N=125)	68.8% (N=275)
mod_strong	1.2576	0.3313	74.00%	31.5% (N=126)	68.5% (N=274)

* Defined as effective bias reduction in the literature
All replications were completed with a caliper width = 0.6

Table 21. Summary of Balance for Simulated Match (N=300)

Table 21. Summary of Balance for Simulated Match (N=300)					
	<u>Before Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>After Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>Balance</u> <u>Improvement</u>	<u>Successfully</u> <u>Matched</u>	<u>Excluded</u>
strong_none	0.0978	0.0070	94.54%*	26.0% (N=78)	74.0% (N=222)
strong_weak	0.3785	0.0444	89.01%	38.7% (N=116)	61.3% (N=184)
strong_mod	0.8292	0.1539	81.79%	36.0% (N=108)	64.0% (N=192)
none_strong	1.2546	0.3139	75.24%	31.0% (N=93)	69.0% (N=207)
weak_strong	1.2585	0.3168	75.06%	31.0% (N=93)	69.0% (N=207)
mod_strong	1.2542	0.3123	75.40%	31.0% (N=93)	69.0% (N=207)

* Defined as effective bias reduction in the literature
All replications were completed with a caliper width = 0.6

Table 22. Summary of Balance for Simulated Match (N=200)

Table 22. Summary of Balance for Simulated Match (N=200)					
	<u>Before Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>After Match</u> <u>Standardized</u> <u>Mean Difference</u>	<u>Balance</u> <u>Improvement</u>	<u>Successfully</u> <u>Matched</u>	<u>Excluded</u>
strong_none	0.1208	0.0081	94.78%*	24.5% (N=49)	75.5% (N=151)
strong_weak	0.3727	0.0393	90.40%	36.0% (N=72)	64.0% (N=128)
strong_mod	0.8612	0.1509	83.12%	34.5% (N=69)	65.5% (N=131)
none_strong	1.2536	0.2996	76.49%	30.5% (N=61)	69.5% (N=139)
weak_strong	1.2697	0.3119	75.75%	30.5% (N=61)	69.5% (N=139)
mod_strong	1.2624	0.3034	76.38%	30.5% (N=61)	69.5% (N=139)

* Defined as effective bias reduction in the literature
All replications were completed with a caliper width = 0.6

The findings for covariate association to outcome and to treatment were again consistent across all conditions of the sample sizes for a reduction in bias. For all of the replications across sample sizes, the mean balance improvement was best for the covariates that were strongly related to outcome. These findings coincide with the recommendation from Rubin and Thomas to include all variables thought to be related to outcome, regardless of the relation to treatment (1996). The covariate relationship magnitude strong_none (strong relationship to DV_no relationship to treatment) displayed the best overall mean bias reduction across sample size conditions. The covariate magnitude selection that had the worst balance improvement for most sample sizes was mod_strong (moderate relationship to DV_strong relationship to treatment). For the smallest sample size of 200, the relationship weak_strong (weak relationship to DV_strong relationship to treatment) was the worst. The findings by Brookhart and colleagues that covariates that have a weak relationship to DV and a strong relationship to treatment (i.e., weak_strong) are confounders that should be eliminated in sample sizes smaller than 500 was found in this study. In addition, covariates that had a mod_strong relationship (moderate to DV_strong to treatment) also increased levels of bias to a similar level. Using the same criterion for what is considered effective bias reduction in the literature (balance improvement above 90%), only the covariate relationship strong_none was able to be deemed effective matching for all sample sizes. The association strong_weak was close to achieving 90% balance improvement across all sample sizes. The figure below further reveals this trend.

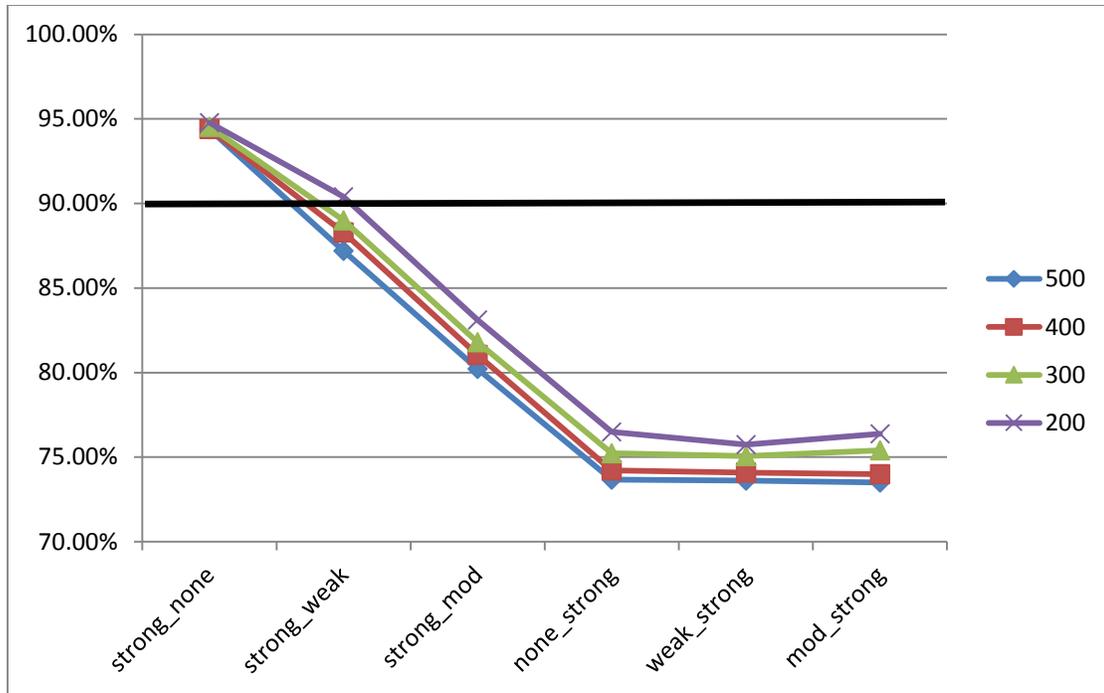


Figure 17. Bias Reduction by Sample Size for Simulation

The mean c-statistic for the simulated dataset was calculated across the conditions for each of the replications to assess the predictive power of the models. This statistic provides information for how effective the model is at predicting which subjects are in the treated group, based on the covariates. The table below displays the mean c-statistic for each of the selected associations for covariates to dependent variable and to treatment across each of the randomly selected sample sizes.

Table 23. Summary of Mean c-statistics by Sample Size for Simulation

DV/T _x	Sample Size			
	500	400	300	200
strong_none	0.7688*	0.7699*	0.7661*	0.7703*
strong_weak	0.7806*	0.7806*	0.7795*	0.7749*
strong_mod	0.7724*	0.7718*	0.7747*	0.7725*
none_strong	0.6407	0.6384	0.6395	0.6424
weak_strong	0.6813	0.6847	0.6823	0.6842
mod_strong	0.7272*	0.7230*	0.7261*	0.7289*

* Defined as favorable predictive power in the literature

**Defined as strong predictive power in the literature

All replications were completed with a caliper width = 0.6

The mean c-statistic is fairly stable across all randomly selected sample sizes. The figure below displays the overall trend for each covariate relationship.

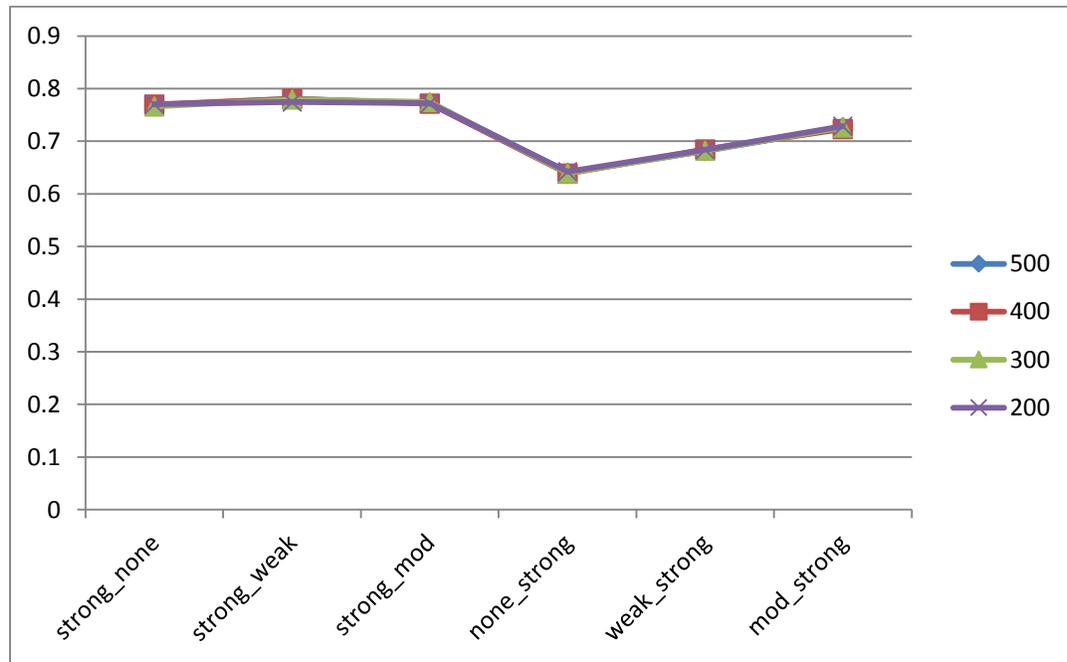


Figure 18. Summary of Mean c-statistics by Sample Size for Simulation

The c-statistics ranged from 0.64 to 0.78, staying consistent across the reductions in sample size. The simulation with the association strong_weak had the highest predictive power, while the model with the association none_strong had the weakest. These findings were true for all sample sizes selected. Using the criteria set forth by Hosmer and colleagues (2000) that a c-

statistic higher than 0.70 is considered favorable and greater than 0.80 strong, all covariate selections would be deemed favorable except for the relationship none_strong and weak_strong. The c-statistics for these magnitudes were just under favorable across the sample sizes selected, with the average being 0.64 and 0.68 respectively. These findings suggest that covariates that are not associated or only weakly associated to the dependent variable could decrease the predictive power of the matching model, regardless of sample size. Further covariates that are strongly related to the dependent variable have high predictive power in the model, regardless of association to the treatment or sample size. Covariates that are strongly related to the treatment require at least a moderate association to the dependent variable to be deemed favorably predictive, regardless of sample size.

A further analyses of the covariate relationship to outcome and treatment was evaluated the bias in estimating the “true” treatment effect for matched and unmatched groups. The figure below displays the findings for all sample sizes and covariate associations with the caliper width of 0.6.

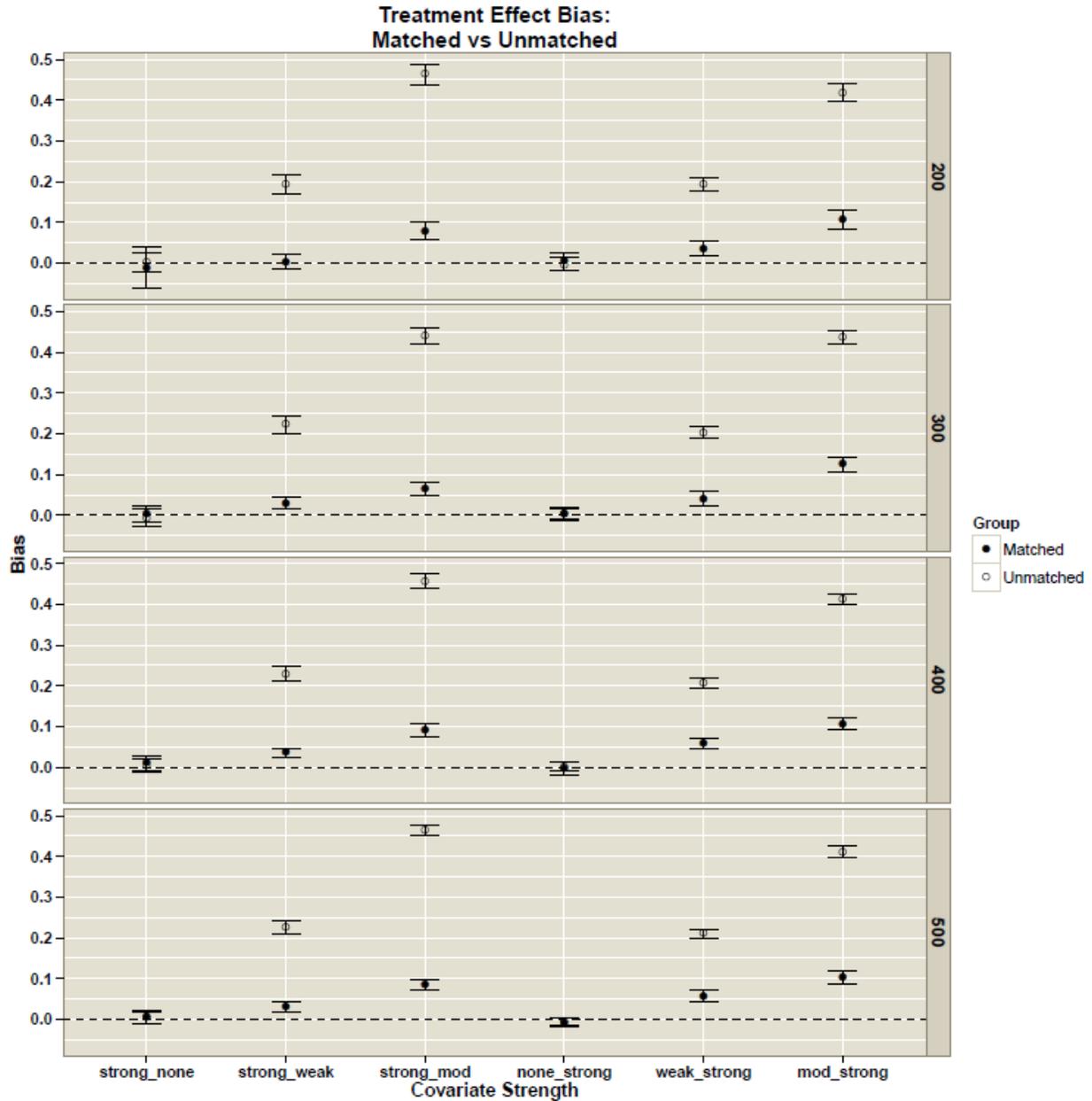


Figure 19. Bias in Estimated Treatment Effect for Matched and Unmatched

Findings for estimating the “true” treatment effect are again consistent across sample sizes for reduction in bias. These results display that the relationships strong_none and none_strong estimate the treatment effect the closest and without bias at the extremes regardless of being matched, as expected. The relationships that displayed the greatest reductions in bias

when estimating the treatment effect after the match (as compared to prior to being matched) were strong_moderate and moderate_strong across all sample sizes. The covariate associations strong_weak and weak_strong displayed similar, smaller reductions in bias when estimating the treatment effect after matching.

Chapter Five – Discussion

Overall, limitations in sample size were assessed as it related to covariate selection and caliper width. The original dataset served as a baseline or “true score” for best possible bias reduction. Sample sizes were randomly reduced from this original dataset to assess bias reduction as it related to caliper width. A Monte Carlo simulation was created based on the original dataset to assess covariate selection as it related to the strength of the relationship to the treatment and to the outcome with the same limitations in sample size and a set caliper width of 0.6. These analyses were replicated 100 times in the R software program to increase power and precision of the findings. A complete copy of the R code used to complete this analysis is available in Appendix C.

Sample Size

Overall, the findings for sample size reduction are surprising. Contrary to the hypothesis that with reduction in sample size the balance improvement would drop below what is considered effective bias reduction, the reduction in bias (or balance improvement) ranged from 96.77% to 99.93% for the sample sizes selected. In the literature, effective bias reduction in the standardized mean differences less than ten percent after matching was considered effective. For this particular sample, randomly reducing the sample to even 200 didn't produce matches that were considered ineffective for reduction in bias with the varying caliper widths.

Caliper Width

The overall findings for all of the 16 conditions were expected for calipers. A narrower caliper width resulted in reducing bias by reducing the systematic differences in the treated and

control groups, but also reduced the final matched sample size. Using a wider caliper width retained more subjects, but increased the bias and systematic differences in the two groups with less overall balance improvement in the overall design. As discussed previously with sample sizes that are limited, the increased variance in the estimated treatment effect will affect the precision of the estimated treatment effect.

The size of the caliper is determined a priori by the investigator. Most researchers use a caliper width of 0.2 standard deviations of the logit of the propensity score due to research suggesting that this caliper size would reduce 99% of the bias in propensity scores (Austin, 2008, 2009; Austin et al., 2007) while others have used a caliper width of 0.6 as this same research stated this caliper width would reduce 90% of the bias (Ayanian et al., 2002). The findings of this research found that with sample sizes less than 500 the most commonly used caliper width of 0.2 effectively reduced the bias consistently around 99.7%, confirming the previous findings by Austin and his colleagues. However, in this sample the caliper width of 0.6 consistently reduced bias by at least 96.77%. These findings show greater reduction in bias with this wider caliper width than that was previously noted by Ayanian and colleagues in 2002. At no point, over the 16 conditions selected for this study did the reduction in bias fall below the ten percent reduction that was deemed as ineffective matching from the literature review. This suggests that with limited sample sizes below 500, researchers could consider a wider caliper in order to retain more subjects and still be able to stay within what is deemed as effectively removing bias in the matched sample, when using the same matching criterion that was utilized in this study. A notable finding is that caliper widths provided similar levels of bias reduction across all sample size conditions until the widest sample size of 0.6. At this point, the bias reduction varies from 98.17% for a sample size of 500 to 96.77% for a sample size of 200. Although each caliper width

provided balance improvement at acceptable levels, the widest caliper one could chose before much dispersion would be 0.3. This suggests that researchers could consider the minimum caliper width of 0.3 when limited by sample size, as they performed similarly.

Covariate Selection

Propensity score matching collapses all covariates into one predictive scalar, regardless of the strength of the association to outcome or treatment. This analysis was completed to determine the optimal covariate relationships to outcome and to treatment for limited sample sizes. For all of the replications across sample sizes, the mean balance improvement was best for the covariate relationship magnitude strong_none (strong relationship to DV_no relationship to treatment). The covariate magnitude selection that had the worst balance improvement for most sample sizes was mod_strong (moderate relationship to DV_strong relationship to treatment). For the smallest sample size of 200, the relationship weak_strong (weak relationship to DV_strong relationship to treatment) was the worst. The same findings by Brookhart and colleagues that covariates that have a weak relationship to DV and a strong relationship to treatment (i.e., weak_strong) are confounders that should be eliminated in sample sizes smaller than 500 was also found in this study. In addition, covariates that had a mod_strong relationship (moderate to DV_strong to treatment) also increased levels of bias to a similar level yet the c-statistic showed favorable predictive power. Using the same criterion for what is considered effective bias reduction in the literature (balance improvement above 90%), only the covariate relationship strong_none was able to be deemed effective matching for all sample sizes. The association strong_weak was close to achieving 90% balance improvement across all sample sizes. These findings suggest that ideal covariates are those that are strongly related to the

outcome variable and only weakly or moderately related to treatment when sample sizes are limited. Inclusion of covariates that are only slightly predictive of the outcome should be used with caution as they can reduce the efficiency of the relevant covariates.

Limitations and Future Research

There are certain limitations to the current study. The main objective of this study was to evaluate the effectiveness of propensity score matching at reducing bias with limited sample size both in a real dataset and in an ideal scenario created using a Monte Carlo simulation study. The current study only examined the matching methodology using nearest neighbor or greedy matching. There is also optimal matching in which the objective is to find the matched pairs with the smallest average distance across all pairs. The two options were found to be comparable in terms of producing a balanced matched sample, yet nearest neighbor is the prevailing approach. Optimal matching uses the smallest average distance across all pairs and does not use caliper widths. Therefore, the findings of this study regarding caliper widths are not applicable to optimal matching. Further, the current study conducted matching using one-to-one without replacement. There are also the matching methods many-to-one or many-to-many. These methods were not discussed as they are rarely used. Future research should include all matching methods.

A possible limitation in the current study is in the basic method of creating the matched groups. This study used logistic regression for creating the propensity score as it is the most commonly used approach. There are other possible methods for calculating a propensity score such as the probit model, discriminant function analysis (DFA) and boosted regression trees.

Future research should include other methods of calculating a propensity score when analyzing the possible effects of limited sample size.

The sample sizes and caliper widths selected for the current study were chosen based on a literature review of information that is lacking or frequency of use. Findings indicate that the threshold for what is deemed acceptable bias reduction lies in the caliper width and not necessarily in limitations in sample size. However, future research should find the true breaking point for both sample size and caliper width instead of preselected categories based on bias reduction. Minimum sample sizes less than 100 were originally conditions for this study, however, sample sizes less than 200 would not converge. Sample sizes were iteratively tried on for this particular sample and convergence was successful with approximately 185 subjects. Further, caliper widths 0.1, 0.2, and 0.3 behaved similarly for all sample size conditions for this dataset until the width of 0.6. Although the amount of bias reduction was acceptable for a caliper width of 0.6 for all sample sizes, future research should locate the exact caliper width between 0.3 and 0.6 in which this dispersion of bias reduction amongst sample sizes first occurs. Further, research on the exact breaking point at which caliper widths become too wide to successfully reduce bias should be conducted.

References

- Althausen, R. P., & Rubin, D. (1970). The Computerized Construction of a Matched Sample. *Amer J Sociol*, 76(2), 325-346.
- Angrist, J., & Hahn, J. (2004). When to control for covariates? Panel asymptotics for estimates of treatment effects. *Review of Economics and statistics*, 86(1), 58-72.
- Augurzky, B., & Schmidt, C. M. (2001). The propensity score: A means to an end.
- Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12), 2037-2049.
- Austin, P. C. (2009). Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations. *Biometrical Journal*, 51(1), 171-184.
- 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(4), 734-753.
- Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in medicine*, 25(12), 2084-2106.
- Ayanian, J. Z., Landrum, M. B., Guadagnoli, E., & Gaccione, P. (2002). Specialty of ambulatory care physicians and mortality among elderly patients after myocardial infarction. *New England Journal of Medicine*, 347(21), 1678-1686.
- Bai, H. (2011). Using propensity score analysis for making causal claims in research articles. *Educational Psychology Review*, 23(2), 273-278.
- Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1980). *Issues in the analysis of selectivity bias*: University of Wisconsin, Inst. for Research on Poverty.
- Black, D. A., & Smith, J. A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of econometrics*, 121(1), 99-124.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, 163(12), 1149-1156.
- Bryson, A., Dorsett, R., & Purdon, S. (2002). The use of propensity score matching in the evaluation of active labour market policies.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1), 31-72.
- Christakis, N. A., & Iwashyna, T. J. (2003). The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social science & medicine*, 57(3), 465-475.
- Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.
- Cole, J. A., Loughlin, J. E., Ajene, A. N., Rosenberg, D. M., Cook, S. F., & Walker, A. M. (2002). The effect of zanamivir treatment on influenza complications: a retrospective cohort study. *Clinical therapeutics*, 24(11), 1824-1839.
- d'Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, 17(19), 2265-2281.

- Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151-161.
- Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49(4), 1231-1236.
- Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly*, 55(1), 74-79.
- Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1), 21-29.
- Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1), 77-90.
- Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589(1), 63-93.
- Glynn, R. J., Schneeweiss, S., & Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology*, 98(3), 253-259.
- Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3), 300-306.
- Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4), 405-420.
- Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28(4), 357-383.
- Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (Vol. 12): Sage Publications.
- Hall, J. A., Summers, K. H., & Obenchain, R. L. (2003). Cost and utilization comparisons among propensity score-matched insulin lispro and regular insulin users. *Journal of Managed Care Pharmacy*, 9(3), 263-268.
- Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*: Springer.
- Haviland, A., Nagin, D. S., Rosenbaum, P. R., & Tremblay, R. E. (2008). Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. *Developmental psychology*, 44(2), 422.
- Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data: National bureau of economic research.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.
- Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology*, 35(1), 1-97.
- Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4), 259-278.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2004). MatchIt: Matching Software for Causal Inference. *Version 0.8. Used with permission*.

- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3), 199-236.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). Assessing the fit of the model. *Applied Logistic Regression, Third Edition*, 153-225.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2), 481-502.
- Imbens, G. M., & Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation: National Bureau of Economic Research.
- Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology*, 150(4), 327-333.
- Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American journal of epidemiology*, 163(3), 262-270.
- Lane, F. C., To, Y. M., Shelley, K., & Henson, R. K. (2012). An illustrative example of propensity score matching with education research. *Career and Technical Education Research*, 37(3), 187-212.
- Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in east germany after unification. *Journal of Business & Economic Statistics*, 17(1), 74-90.
- Luellen, J. K., Shadish, W. R., & Clark, M. (2005). Propensity Scores An Introduction and Experimental Test. *Evaluation Review*, 29(6), 530-558.
- Magee, M. J., Coombs, L. P., Peterson, E. D., & Mack, M. J. (2003). Patient selection and current practice strategy for off-pump coronary artery bypass surgery. *Circulation*, 108(10 suppl 1), II-9-II-14.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2, 49-55.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (Vol. 1): Psychology Press.
- McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4), 403.
- Moss, R. R., Humphries, K. H., Gao, M., Thompson, C. R., Abel, J. G., Fradet, G., & Munt, B. I. (2003). Outcome of mitral valve repair or replacement: a comparison by propensity score analysis. *Circulation*, 108(10 suppl 1), II-90-II-97.
- Murray, P. K., Singer, M., Dawson, N. V., Thomas, C. L., & Cebul, R. D. (2003). Outcomes of rehabilitation services for nursing home residents. *Archives of physical medicine and rehabilitation*, 84(8), 1129-1136.
- Neyman, J., & Iwazskiewicz, K. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 107-180.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96-146.
- Pearl, J. (2011). Invited commentary: understanding bias amplification. *American journal of epidemiology*, 174(11), 1223-1227.
- Peikes, D. N., Moreno, L., & Orzol, S. M. (2008). Propensity score matching. *The American Statistician*, 62(3).

- Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X. H., & Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety*, 9(2), 93-101.
- Ridgeway, G., McCaffrey, D. F., Morral, A. R., Burgette, L. F., & Griffin, B. A. (2013). Toolkit for weighting and analysis of nonequivalent groups.
- Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 479-495.
- Rosenbaum, P. R. (2002). Sensitivity to hidden bias *Observational studies* (pp. 105-170): Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 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(387), 516-524.
- Rosenbaum, P. R., & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, 41(1), 103-116.
- Rosenbaum, P. R., & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33-38.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.
- Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, 109-120.
- Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics*, 36(2), 293-298.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2), 757-763.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4), 169-188.
- Rubin, D. B. (2006). *Matched sampling for causal effects*: Cambridge University Press.
- Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 249-264.
- Schutt, R. K. (2011). *Investigating the social world: The process and practice of research*: Pine Forge Press.
- Seeger, J. D., Walker, A. M., Williams, P. L., Saperia, G. M., & Sacks, F. M. (2003). A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. *The American journal of cardiology*, 92(12), 1447-1451.
- Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6), 546-555.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*: Wadsworth Cengage learning.
- Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1), 305-353.

- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1), 1.
- Stuart, E. A., Marcus, S. M., Horvitz-Lennon, M. V., Gibbons, R. D., & Normand, S.-L. T. (2009). Using non-experimental data to estimate treatment effects. *Psychiatric annals*, 39(7), 414-51.
- Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics.
- Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and drug safety*, 13(12), 841-853.
- Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and program planning*, 28(2), 209-220.
- Yu, D. T., Platt, R., Lankester, P. N., Black, E., Sands, K. E., Schwartz, J. S., . . . Snyderman, D. R. (2003). Relationship of pulmonary artery catheter use to mortality and resource utilization in patients with severe sepsis*. *Critical care medicine*, 31(12), 2734-2741.
- Zhao, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and statistics*, 86(1), 91-107.

Appendix A

Data Use Agreement between Kansas City Kansas Public Schools And Stephani Howarter at the University of Kansas

This Data Use Agreement, effective on November 3, 2014, by and between the Kansas City Kansas Public Schools (hereafter "KCKPS") and a Stephani Howarter (hereafter "Recipient") at the University of Kansas on behalf of the University of Kansas Center for Research, Inc.

1. This agreement sets forth the terms and conditions pursuant to which KCKPS will disclose the limited data set from the previously collected study #19741 (under the Data Agreement between KCKPS and KU-CEOP) to Recipient.
2. **Individuals Who May Access the Data:** Except as otherwise specified herein, Recipient may make all uses and disclosures of the Limited Data Set necessary to conduct the research described herein: KU GEAR UP Comparative Study for Dissertation Research from previously collected data study #19741. Stephani Howarter, doctoral candidate at KU are permitted to receive and use this dataset for purposes of this Agreement.

In addition to Recipient, the following individuals, or classes of individuals, are permitted to use or receive the Limited Data Set for purposes of the research project: NONE

3. **Purpose:** In order to evaluate the implementation and outcomes of the college readiness services, provided by GEAR UP, Stephani Howarter will require access to the student-level demographics and academic data for JC Harmon, Washington, Schlagle, and Wyandotte High Schools. This data has been previously collected in study #19741. Ms. Howarter would only require access to this previously collected data in order to further analyze this dataset with different statistical methodology and inform the current KU GEAR UP evaluation as well as model a "real data" example for a statistical methodology being analyzed in a dissertation.

The following data already provided by KCKPS to KU-CEOP will be permitted to Recipient:

- Student-level demographics, educational programming and services
- High School grade point average data
- EXPLORE, PLAN, and ACT scores

*All data would have any identifying information removed prior to Ms. Howarter's access.

4. **Permitted Uses and Disclosures:** Recipient agrees to use appropriate safeguards to protect the data from misuse or inappropriate disclosure and to prevent use or disclosure of the Limited Data Set other than as provided for by this DUA or as otherwise required by law or regulation.
5. Recipient will not use or disclose the data for any purpose other than permitted by this Agreement pertaining to this project, or as required by law. If disclosure of data of any kind is deemed necessary, it will take place only after prior notification to KCKPS. Recipient agrees to

ensure that any agent, including a subcontractor, to whom he or she provides the Limited Data Set, agrees to the same restrictions and conditions that apply through this DUA, with respect to such information.

6. Recipient shall not attempt to identify the individuals to whom the data pertains in any report or material, or attempt to contact such individuals.
7. This DUA shall be effective on the Effective Date set forth above and shall continue until this Agreement expires, June 1, 2015, unless extended by mutual agreement in writing as a modification to this agreement. Recipient may terminate this Agreement by returning or destroying the data and providing written notice thereof to Covered Entity. All data connected with this project shall be destroyed when no longer needed for the purposes for which the Project was conducted.

[Kansas City Kansas Public Schools]

Print Name DAVID RAND

Signature: 

Title: DIRECTOR DERA

[Stephani Howarter, Recipient]

Print Name: Stephani Howarter Stephani Howarter

Signature: 

Title Doctoral Candidate

Data Use Agreement between
Kansas City Kansas Public Schools
And
Stephani Howarter at the University of Kansas

This Data Use Agreement, effective on November 3, 2014, by and between the Kansas City Kansas Public Schools (hereafter "KCKPS") and a Stephani Howarter (hereafter "recipient") at the University of Kansas on behalf of the University of Kansas Center for Research, Inc.

Modification:

This DUA shall be effective on the Effective Date set forth above and shall continue until this agreement expires, December 31, 2015, as extended by mutual agreement in writing as a modification to the previous data use agreement. Recipient may terminate this Agreement by returning or destroying the data and providing written notice thereof to Covered Entity. All data connected with this project shall be destroyed when no longer needed for the purposes for which the Project was conducted.

[Kansas City Kansas Public Schools]

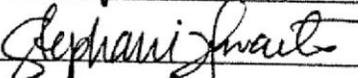
Print Name DAVID RAND

Signature 

Title DIRECTOR DERA

[Stephani Howarter, Recipient]

Print Name Stephani Howarter

Signature 

Title Doctoral candidate

Appendix B



APPROVAL OF PROTOCOL

June 25, 2015

Stephani Howarter

Dear Stephani Howarter:

On 6/25/2015, the IRB reviewed the following submission:

Type of Review:	Initial Study
Title of Study:	The Efficacy of Propensity Score Matching with Limited Sample Sizes
Investigator:	Stephani Howarter
IRB ID:	STUDY00002633
Funding:	None
Grant ID:	None
Documents Reviewed:	• HSCL Initial Submission Form , • Data Sharing Agreement with KCKPS, • Extension Data Sharing Agreement ,

The IRB approved the study on 6/25/2015.

1. Notify HSCL about any new investigators not named in the original application. Note that new investigators must take the online tutorial at https://rgs.drupal.ku.edu/human_subjects_compliance_training.
2. Any injury to a subject because of the research procedure must be reported immediately.
3. When signed consent documents are required, the primary investigator must retain the signed consent documents for at least three years past completion of the research activity.

Continuing review is not required for this project, however you are required to report any significant changes to the protocol prior to altering the project.

Please note university data security and handling requirements for your project:
<https://documents.ku.edu/policies/IT/DataClassificationandHandlingProceduresGuide.htm>

You must use the final, watermarked version of the consent form, available under the "Documents" tab in eCompliance.

Sincerely,

Stephanie Dyson Elms, MPA
IRB Administrator, KU Lawrence Campus

Appendix C

```
#####Master File PSM Code for "true/full" match data
#####Last edited 20151016

setwd("C:/Users/stephani.howarter/Desktop/Master File PSM/data")
dir()
data1 <- read.table("Master File PSM.csv", header=TRUE, sep=",")
head(data1, 5)

#####calculate descriptive stats for the covariates before matching
#####

#####calculate frequencies for covariates for "nontreated" group
data7 <- data1[data1$GROUP== 0,]
table(data7$Grade)
table(data7$Gender)
table(data7$Ethnic)
table(data7$Lunch)
table(data7$SWD)
table(data7$ELL)
rm(data7)
#####

#####calculate frequencies for covariates for "treated" group
data7 <- data1[data1$GROUP== 1,]
table(data7$Grade)
table(data7$Gender)
table(data7$Ethnic)
table(data7$Lunch)
table(data7$SWD)
table(data7$ELL)
rm(data7)
#####

#####end
#####chi-square test and cramer V to show group differences in covariates before matching
#####criteria of cramer V: small, 0.1 medium 0.3, large 0.5
#####

z1=chisq.test(data1$GROUP, data1$Grade, correct=FALSE)
z1
z=chisq.test(data1$GROUP, data1$Grade, correct=FALSE)$statistic
cramer.of.Grade= sqrt(as.numeric(z)/1726)
cramer.of.Grade
# calculate the cramer V between "GROUP before match" and Grade

z1=chisq.test(data1$GROUP, data1$Gender, correct=FALSE)
z1
z=chisq.test(data1$GROUP, data1$Gender, correct=FALSE)$statistic
```

```

cramer.of.Gender= sqrt(as.numeric(z)/1726)
cramer.of.Gender
# calculate the cramer V between "GROUP before match" and gender

z1=chisq.test(data1$GROUP, data1$Ethnic, correct=FALSE)
z1
z=chisq.test(data1$GROUP, data1$Ethnic, correct=FALSE)$statistic
cramer.of.Ethnic= sqrt(as.numeric(z)/1726)
cramer.of.Ethnic
# calculate the cramer V between "GROUP before match" and Ethnic

z1=chisq.test(data1$GROUP, data1$Lunch, correct=FALSE)
z1
z=chisq.test(data1$GROUP, data1$Lunch, correct=FALSE)$statistic
cramer.of.Lunch= sqrt(as.numeric(z)/1726)
cramer.of.Lunch
# calculate the cramer V between "GROUP before match" and Lunch

z1=chisq.test(data1$GROUP, data1$SWD, correct=FALSE)
z1
z=chisq.test(data1$GROUP, data1$SWD, correct=FALSE)$statistic
cramer.of.SWD= sqrt(as.numeric(z)/1726)
cramer.of.SWD
# calculate the cramer V between "GROUP before match" and SWD

z1=chisq.test(data1$GROUP, data1$ELL, correct=FALSE)
z1
z=chisq.test(data1$GROUP, data1$ELL, correct=FALSE)$statistic
cramer.of.ELL= sqrt(as.numeric(z)/1726)
cramer.of.ELL
# calculate the cramer V between "GROUP before match" and ELL
#####end

#####

##### library the packages need for PSM
library(MASS)
library(digest)
library("MatchIt")
library(VIM)
#####

#####create the criteria for distance for PSM, CaliperValue =.2
#####
glmResult <- glm(GROUP ~ Grade + Gender + Ethnic + Lunch +SWD + ELL, binomial,
data= data1)

a= predict(glmResult,data1, type="response")
b= 1-a
LogistScore= log((a/b),10)
CaliperValue=sqrt(var(LogistScore))*0.2

#####

```

```

#####MatchIt for PSM
#####
m.out <-matchit(GROUP ~ Grade + Gender + Ethnic + Lunch + SWD + ELL, data= data1, method= "nearest", caliper=
CaliperValue, distance= "logit", m.order="largest", replace = FALSE, ratio=1)

#### GROUP is the membership variable (1= treatment group, 0= control group)
#### X is the covariate variables used to match the data
#### method= nearest neighbor matching technique. Default method of MatchItpackage.
#### caliper value should be set to be 0.2* S.D of propensity score according to literature calculated as CaliperValue
#### ratio = 1 means 1:1 matching; ratio = 2 means 2 nontreated matched to one treated

#####report PSM results, and put result into the file call match.data1
m.out
summary(m.out)
match.data1 <- match.data(m.out)

#### Provides which T matched to which C
m.outmatrix <- m.out$match.matrix # This gives us the matched matrix
m.outmatrix

#####will provide QQ plot, jitter plot, and histogram

plot(m.out)
plot(m.out, type = "jitter")
plot(m.out, type="hist")
#####

#####extract matched data set & check; writes csv file "matched" to folder
match.data1 <- match.data(m.out)

## check that it has the expect number of rows
nrow(match.data1)

## export the data to a CSV file
write.table(match.data1, file = "matched.csv", sep = ",", row.names = FALSE)

#####

#####install package to display C statistics before and after matching

library(rms)
#####

#####displays c stat before matched
lrm(formula = GROUP ~ Grade + Gender + Ethnic + Lunch +SWD + ELL, data=data1)

#####displays c stat after matched; successful match = drop in c stat
lrm(formula = GROUP ~ Grade + Gender + Ethnic + Lunch +SWD + ELL, data= match.data1)

#####

```

```

#####calculate the descriptive stats for covariates after matching
#####

#####calculate freq for covariates for nontreated group after match
data7 <- match.data1[match.data1$GROUP== 0,]
table(data7$Grade)
table(data7$Gender)
table(data7$Ethnic)
table(data7$Lunch)
table(data7$SWD)
table(data7$ELL)
rm(data7)
#####

#####calculate freq for covariates for treated group after match
data7 <- match.data1[match.data1$GROUP== 1,]
table(data7$Grade)
table(data7$Gender)
table(data7$Ethnic)
table(data7$Lunch)
table(data7$SWD)
table(data7$ELL)
rm(data7)
#####

#####use the cramer V to show that group difference in covariates are small after matching
#####criteria of cramer V: small, 0.1 medium 0.3, large 0.5 (Cohen)
#####

z1=chisq.test(match.data1$GROUP, match.data1$Grade, correct=FALSE)
z1
z=chisq.test(match.data1$GROUP, match.data1$Grade, correct=FALSE)$statistic
cramer.of.Grade= sqrt(as.numeric(z)/1478)
cramer.of.Grade
# calculate the cramer V between "GROUP after match" and Grade

z1=chisq.test(match.data1$GROUP, match.data1$Gender, correct=FALSE)
z1
z=chisq.test(match.data1$GROUP, match.data1$Gender, correct=FALSE)$statistic
cramer.of.Gender= sqrt(as.numeric(z)/1478)
cramer.of.Gender
# calculate the cramer V between "GROUP after match" and gender

z1=chisq.test(match.data1$GROUP, match.data1$Ethnic, correct=FALSE)
z1
z=chisq.test(match.data1$GROUP, match.data1$Ethnic, correct=FALSE)$statistic
cramer.of.Ethnic= sqrt(as.numeric(z)/1478)
cramer.of.Ethnic
# calculate the cramer V between "GROUP after match" and Ethnic

z1=chisq.test(match.data1$GROUP, match.data1$Lunch, correct=FALSE)

```

```

z1
z=chisq.test(match.data1$GROUP, match.data1$Lunch, correct=FALSE)$statistic
cramer.of.Lunch= sqrt(as.numeric(z)/1478)
cramer.of.Lunch
# calculate the cramer V between "GROUP after match" and Lunch

z1=chisq.test(match.data1$GROUP, match.data1$SWD, correct=FALSE)
z1
z=chisq.test(match.data1$GROUP, match.data1$SWD, correct=FALSE)$statistic
cramer.of.SWD= sqrt(as.numeric(z)/1478)
cramer.of.SWD
# calculate the cramer V between "GROUP after match" and SWD

z1=chisq.test(match.data1$GROUP, match.data1$ELL, correct=FALSE)
z1
z=chisq.test(match.data1$GROUP, match.data1$ELL, correct=FALSE)$statistic
cramer.of.ELL= sqrt(as.numeric(z)/1478)
cramer.of.ELL
# calculate the cramer V between "GROUP after match" and ELL

#####

#### Q1Q2 Howarter
#### Last edited 20151029

#### Clear the workspace
rm(list = ls())

getwd()
setwd("C:/Users/stephani.howarter/Desktop/Master File PSM/data")
dir()
dat <- read.table("Master File PSM.csv", header=TRUE, sep=",")
str(dat)
head(dat)
colnames(dat)
summary(dat)

#### Load required libraries
library(MASS)
library(digest)
library(MatchIt)
library(rms)
library(rockchalk)
#####
#### Function that creates the criteria for distance for PSM with caliper=0.2
#### will compute the c-statistics from original and matched
#### Note: dat is the original data
####browser() to step by step view pulls
createPSMcrit <- function(dat, set_caliper){

####browser()
glmResult <- glm(GROUP ~ Grade + Gender + Ethnic + Lunch +SWD + ELL,

```

```

        family = binomial, data= dat)

#### Compute a, b, LogistScore, and CaliperValue
a <- predict(glmResult, dat, type = "response")
b <- 1 - a
LogistScore <- log((a/b), 10)
CaliperValue <- sqrt(var(LogistScore)) * set_caliper

#### Call matchit to do the PSM

m.out <- matchit(GROUP ~ Grade + Gender + Ethnic + Lunch + SWD + ELL,
                data = dat, method = 'nearest', caliper = CaliperValue,
                distance = 'logit', ratio = 1, replace = FALSE)

#### Saves the summary of m.out as an object
get_m.out_summary <- summary(m.out, standardize = TRUE)

#### Get the data from sum.out

bias_std<- unlist(c(get_m.out_summary$sum.all[1,4, drop=TRUE],
                  get_m.out_summary$sum.matched[1, 1:4, drop = TRUE],
                  get_m.out_summary$reduction[1,1, drop=TRUE],
                  get_m.out_summary$nn[2,1, drop =TRUE]))

#### Get the output data set
m.out_data <- match.data(m.out)

#### Compare matched and unmatched data
unmatched <- lrm(GROUP ~ Grade + Gender + Ethnic + Lunch + SWD + ELL, data = dat)
matched <- lrm(GROUP ~ Grade + Gender + Ethnic + Lunch + SWD + ELL, data = m.out_data)

#### Get the C statistics from matched and unmatched
c_unmatched <- unmatched$stats['C']
c_matched <- matched$stats['C']

#### Return the difference
####out_list <- list(diff, bias_std)
out_list <- list(c_matched, bias_std)
}
#####
##
#### Function runs the simulation
runQ1Q3simulation <- function(row, input_data, cond_mat, n_reps){
  SS_use <- cond_mat[row, 1]
  caliper_use <- cond_mat[row, 2]

  #### Create c-dif storage vector
  ####c_dif_out <- rep(NA, length = n_reps)
  c_matched_out <- rep(NA, length = n_reps)

  #### Create standardized bias list of matrices

```

```

bias_mat <- matrix(NA, nrow = n_reps, ncol = 7)
colnames(bias_mat) <- c("Mean Diff", "Means Treated", "Means Control", "SD Control", "Mean Diff", "Mean Diff.",
"Matched")

#### For loop over the number of replications
for(i in 1:n_reps){
  #### Get pick index
  pick_index <- sample(x = 1:nrow(input_data), size = SS_use, replace = FALSE)

  #### Get a random selection of rows
  ran_samp <- input_data[pick_index, ]

  #### Run the createPSMcrit function
  out_list <- createPSMcrit(dat = ran_samp, set_caliper = caliper_use)

  #### Note first element of out_list corresponds to c-dif and second
  #### element corresponds to the bias matrix
  ####c_dif_out[i] <- out_list[[1]]

  c_matched_out[i] <- out_list[[1]]
  bias_mat[i, ] <- out_list[[2]]

}

#####
##
#### End for loop over replications

#### Return a list
####list(c_dif_out, bias_mat)
list(c_matched_out, bias_mat)
}
####Total number of rows to sample without replacement
totalN <-nrow(dat)

####Sample sizes to sample from
sampSizes <- c(500,400,300,200)

####Create List of caliper values
caliper_vector <- c(0.1,0.2,0.3,0.6)

####Create conditions matrix using Sampsizes and calip_vector
condsMat <- expand.grid(SS = sampSizes, caliper = caliper_vector)
####Number of times to replicate
nReps <- 100

set.seed (062103)

big_list_out <- lapply(1:nrow(condsMat), runQ1Q3simulation,
  input_data = dat,
  cond_mat = condsMat,
  n_reps = nReps)

```

```

#####Note elements of the list correspond to condsMat

##### Removes scientific notation
options(scipen=999)

##### To get mean of the c stats
c_stat_list <- list()
for(i in 1:length(big_list_out)){
  c_stat_list[[i]] <- big_list_out[[i]][[1]]
}
mean_c_stats <- lapply(c_stat_list, mean)

##### to make a vector use unlist(mean_c_stats)
cbind(condsMat, unlist(mean_c_stats))

##### To get means of bias mat
bias_mat_list <- list()
for(i in 1:length(big_list_out)){
  bias_mat_list[[i]] <- big_list_out[[i]][[2]]
}
means_bias_mat <- lapply(bias_mat_list, colMeans)

means_bias_mat

#####

##### Q3 Howarter
##### Last edited 201511101

## Clear the workspace
rm(list = ls())

setwd("C:/Users/stephani.howarter/Desktop/Master File PSM/data")
dir()

library(MASS)
library(digest)
library(mvtnorm)
library(MatchIt)
library(rms)

#####plot packages
#####library(ggplot2)
#####library(tidyr)
#####library(dplyr)
#####library(plyr)

#####

#####

#####

## Function to create vector with the necessary relationships for covar to Treatment and DV

```

```

createCovRelations <- function(covStr){
  ## Series of if else statements to get the conditions
  if(covStr == "strong_none"){

    out <- c(Coef = .6, OR = 1)

  } else if(covStr == "strong_weak"){

    out <- c(Coef = .6, OR = 1.44)

  } else if(covStr == "strong_mod"){

    out <- c(Coef = .6, OR = 2.47)

  } else if(covStr == "none_strong"){

    out <- c(Coef = 0, OR = 4.25)

  } else if(covStr == "weak_strong"){

    out <- c(Coef = .2, OR = 4.25)

  } else if(covStr == "mod_strong"){

    out <- c(Coef = .4, OR = 4.25)

  } else {

    stop("You used something that you should not have")

  }

  cov_vector <- out
  cov_vector
}

## Function to makeData. The function takes the cov_vector from
## createCovRelations function, the sample size, and the treatment
## effect (the default value is 0.5, true treatment effect = .11).
## The function returns a data frame with y, x, and treat for use in
## createPSMcrit function
makeData <- function(cov_vector, sampSize, trt_effect = .11){
  require(mvtnorm)
  Coef <- cov_vector["Coef"]
  OR <- cov_vector["OR"]
  ## Generate x from standard normal
  x <- rnorm(sampSize)
  ## propensity score stuff
  f <- log(OR) * x
  probs <- exp(f)/(1 + exp(f))
  treat <- rbinom(sampSize, 1, probs)
  ## Generate the y values

```

```

y <- Coef * x + trt_effect * treat + rnorm(sampSize, 0, .5)
## Create data frame to output
out_df <- data.frame(y = y, x = x, treat = treat)
attr(out_df, "trt_effect") <- trt_effect
out_df
}

## A function that creates the criteria for distance for PSM with
## caliper value = 0.6, this function will compute the c-statistics from
## original data and matched data
## Note dat is the original data (a single data frame)
createPSMcrit <- function(dat, set_caliper){
  ## Get the treatment effect from the dat attributes
  trt_effect <- attributes(dat)$trt_effect
  ## Put the formula you will use on all data sets along with the normal
  ## glmResult stuff in your file
  glmResult <- glm(treat ~ x, family = binomial, data = dat)

  ## Compute a, b, LogistScore, and CaliperValue as per your syntax file
  a <- predict(glmResult, dat, type = "response")
  b <- 1 - a
  LogistScore <- log((a/b), 10)
  CaliperValue <- sqrt(var(LogistScore)) * set_caliper

  ## call matchit to do the PSM
  m.out <- matchit(treat ~ x, data = dat,
    method = 'nearest', caliper = CaliperValue,
    distance = 'logit', ratio = 1, replace = FALSE)

  ##### Saves the summary of m.out as an object
  get_m.out_summary <- summary(m.out, standardize = TRUE)

  ##### Get the data from sum.out
  ##

  bias_std <- unlist(c(get_m.out_summary$sum.all[1,4, drop=TRUE],
    get_m.out_summary$sum.matched[1, 1:4, drop = TRUE],
    get_m.out_summary$reduction[1,1, drop=TRUE],
    get_m.out_summary$nn[2,1, drop =TRUE]))

  ## Get the output data set
  m.out_data <- match.data(m.out)

  ## compare matched and unmatched data
  unmatched <- lrm(y ~ x + treat, data = dat)
  matched <- lrm(y ~ x + treat, data = m.out_data)

  ## Get the C statistics from matched and unmatched
  c_unmatched <- unmatched$stats['C']
  c_matched <- matched$stats['C']

```

```

out_list <- list(c_matched, bias_std)
}

## This function runs the simulation
runQ2simulation <- function(row, cond_mat, n_reps, caliper_val){
  sampSize <- cond_mat[row, "SS"]
  covStr <- cond_mat[row, "covStr"]
  caliper_use <- caliper_val

  ## Create c-dif storage vector
  ###c_dif_out <- matrix(NA, nrow = n_reps, ncol = 1)
  c_matched_out <- rep(NA, length = n_reps)

  ## Create standardized bias list of matrices

  bias_mat <- matrix(NA, nrow = n_reps, ncol = 7)
  colnames(bias_mat) <- c("Mean Diff", "Means Treated", "Means Control", "SD Control", "Mean Diff", "Mean
Diff.", "Matched")
  ## For loop over the number of replications
  for(i in 1:n_reps){
    print(i)
    ## Get cov_vector
    cov_vector <- createCovRelations(covStr)

    ## make a data set with the appropriate sample size
    ## and covariate stuff
    data_set <- makeData(cov_vector, sampSize, trt_effect = .11)

    ## Run the createPSMcrit function
    out_list <- createPSMcrit(dat = data_set, set_caliper = caliper_use)

    c_matched_out[i] <- out_list[[1]]
    bias_mat[i, ] <- out_list[[2]]

  } ## End for loop over replications
  print(paste0("Finished condition ", row))
  c_matched_out <- suppressWarnings(data.frame(c_matched_out, cond_mat[row, ], cond = row))
  bias_mat <- suppressWarnings(data.frame(bias_mat, cond_mat[row, ], cond = row))
  ## Return a list
  list(c_matched_out, bias_mat)
}

## Function used for summarizing the means and SDs, SEs, and CI for plotting
## This function was taken from R cookbook
summarySE <- function(data=NULL, measurevar, groupvars=NULL, na.rm=FALSE,
  conf.interval=.95, .drop=TRUE) {
  library(plyr)

```

```

# New version of length which can handle NA's: if na.rm==T, don't count them
length2 <- function (x, na.rm=FALSE) {
  if (na.rm) sum(!is.na(x))
  else length(x)
}

# This does the summary. For each group's data frame, return a vector with
# N, mean, and sd

datac <- ddpby(data, groupvars, .drop=.drop,
  .fun = function(xx, col) {
    c(N = length2(xx[[col]], na.rm=na.rm),
      mean = mean (xx[[col]], na.rm=na.rm),
      sd = sd (xx[[col]], na.rm=na.rm)
    )
  },
  measurevar
)

# Rename the "mean" column
datac <- rename(datac, c("mean" = measurevar))

datac$se <- datac$sd / sqrt(datac$N) # Calculate standard error of the mean

# Confidence interval multiplier for standard error
# Calculate t-statistic for confidence interval:
# e.g., if conf.interval is .95, use .975 (above/below), and use df=N-1
ciMult <- qt(conf.interval/2 + .5, datac$N-1)
datac$ci <- datac$se * ciMult

return(datac)
}

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

## Create conditions for conditions matrix
sampSize <- c(500,400,300,200)

covRelations <- c("strong_none", "strong_weak", "strong_mod", "none_strong",
  "weak_strong", "mod_strong")
caliper_val <- 0.6
## create conditions matrix
condsMat <- expand.grid(SS = sampSize, covStr = covRelations)

nReps <- 100
set.seed (062103)

big_list_out <- lapply(1:nrow(condsMat), runQ2simulation,
  cond_mat = condsMat,
  n_reps = nReps, caliper_val = caliper_val)

```

```

#### Removes scientific notation
options(scipen=999)

#####
#### To get mean of the c stats
c_stat_list <- list()
for(i in 1:length(big_list_out)){
  c_stat_list[[i]] <- big_list_out[[i]][[1]]
}
c_stat_list

#### To get means of bias mat
bias_mat_list <- list()
for(i in 1:length(big_list_out)){
  bias_mat_list[[i]] <- big_list_out[[i]][[2]]
}

bias_mat_list

write.csv(bias_mat_list, file="Q3.csv")
write.csv(c_stat_list, file="Q3_cstat.csv")

##### Plotting Code Below #####

##### C DIF #####

## Get all c_dif
c_dif_list <- lapply(big_list_out, function(x) x[[1]])

## Make a big data frame by binding all the rows of the data frames
c_dif_df <- do.call(rbind, c_dif_list)

c_dif_df_plot <- summarySE(c_dif_df, measurevar = 'c_dif_out',
  groupvars = c('cond', 'SS', 'covStr'))

c_dif_df_plot <- mutate(c_dif_df_plot, SS = factor(SS, levels = c(200, 300, 400, 500),
  labels = c('200', '300', '400', '500')))

gg_bias <- ggplot(c_dif_df_plot, aes(x = covStr, y = c_dif_out)) +
  geom_errorbar(aes(ymin = c_dif_out - ci, ymax = c_dif_out + ci), width = .2) +
  geom_point(size = 2) +
  geom_hline(yintercept = 0.0, colour='black', linetype = "dashed")

gg_bias <- gg_bias + ggtitle("C Difference Statistic") + theme_bw() +
  theme(panel.background = element_rect(fill = 'gray88'),
  panel.grid.major = element_line(colour = "white", size = .5),
  panel.grid.minor = element_line(colour = "white", size = 0.3)) + ylab("C Difference") +
  xlab(paste0("Covariate Strength")) +
  theme(axis.title.x = element_text(face = "bold"), axis.title.y = element_text(face = "bold"),
  plot.title = element_text(face = "bold"), axis.text.x=element_text(face = "bold"),
  axis.text.y=element_text(face="bold")) +

```

```

facet_grid(SS~.) + theme(strip.text.y = element_text(face = "bold", size = 11), strip.text.x =
  element_text(face = "bold", size = 11))
ggsave("CDF.pdf", width = 10, height = 10, units = "in")

##### Bias list #####

## Get all bias
bias_list <- lapply(big_list_out, function(x) x[[2]])

## Make a big data frame by binding all the rows of the data frames
bias_df <- do.call(rbind, bias_list)

## The following code will be used to put the data frame in long format
## for ggplot2
bias_df_long <- gather(bias_df, match_type, dv, Matched:Unmatched)

bias_df_plot <- summarySE(bias_df_long, measurevar = 'dv',
  groupvars = c('cond', 'SS', 'covStr', 'match_type'))
bias_df_plot <- mutate(bias_df_plot, SS = factor(SS, levels = c(200, 300, 400, 500),
  labels = c('200', '300', '400', '500')))

## Plot bias
gg_bias <- ggplot(bias_df_plot, aes(x = covStr, y = dv, shape = match_type)) +
  geom_errorbar(aes(ymin = dv - ci, ymax = dv + ci), width = .2) +
  geom_point(size = 2) +
  scale_shape_manual(values = c(19, 1), labels = c('Matched', 'Unmatched'), name = "Group") +
  geom_hline(yintercept = 0.0, colour = 'black', linetype = "dashed")

gg_bias <- gg_bias + ggtitle("Treatment Effect Bias:\nMatched vs Unmatched") + theme_bw() +
  theme(panel.background = element_rect(fill = 'gray88'),
    panel.grid.major = element_line(colour = "white", size = .5),
    panel.grid.minor = element_line(colour = "white", size = 0.3)) + ylab("Bias") +
  xlab(paste0("Covariate Strength")) +
  theme(axis.title.x = element_text(face = "bold"), axis.title.y = element_text(face = "bold"),
    plot.title = element_text(face = "bold"), axis.text.x = element_text(face = "bold"),
    axis.text.y = element_text(face = "bold")) +
  facet_grid(SS~.) + theme(strip.text.y = element_text(face = "bold", size = 11), strip.text.x =
    element_text(face = "bold", size = 11))
ggsave("TreatmentEffectBias.pdf", width = 10, height = 10, units = "in")

```