THE PERFORMANCE OF PROPENSITY SCORE METHODS TO ESTIMATE THE AVERAGE TREATMENT EFFECT IN OBSERVATIONAL STUDIES WITH SELECTION BIAS: A MONTE CARLO SIMULATION STUDY

By

SUNGUR GUREL

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN EDUCATION

UNIVERSITY OF FLORIDA

2012

To my family

ACKNOWLEDGMENTS

First of all I would like to thank Dr. Walter Leite and Dr. James Algina for guiding me in my thesis. I thank to faculty and students of Research and Evaluation Methodology Program. I would like to thank Turkish Government for the financial support. I also would like to thank Veysel Duman, Suleyman Tor, and Halit Yilmaz for trusting me. Finally, I would like to thank Dilek C. Gulten for her support from the beginning.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 7

LIST OF FIGURES 8

LIST OF ABBREVIATIONS 9

ABSTRACT 10

CHAPTER

1 INTRODUCTION 12

2 THEORETICAL FRAMEWORK 15

The Potential Outcomes Framework 15
Propensity Score Methods for Reducing Selection Bias in ATE Estimates 17
Inverse Probability of Treatment Weighting (IPTW) 17
Truncated Inverse Probability of Treatment Weighting (TIPTW) 18
Propensity Score Stratification (PSS) 19
Optimal Full Propensity Score Matching (OFPSM) 20
Standard Error Estimation with Propensity Score Methods 21
Comparison of Propensity Score Methods' Performances 24

3 METHOD 27

Data Simulation 27
Estimation of ATE and Standard Errors 27
Analyses 31

4 RESULTS 33

5 DISCUSSION 41

APPENDIX

A RELATIVE BIAS OF ATE ESTIMATES 44

B RELATIVE BIAS OF STANDARD ERRORS OF THE ATE ESTIMATES 49

C EMPRICAL COVERAGE AND POWER TABLES 54

REFERENCES 56

BIOGRAPHICAL SKETCH 61

LIST OF TABLES

Table		page
A-1	Relative bias of ATE estimates in the baseline condition	44
A-2	Relative bias of ATE estimates and the percent bias reduction of ATE estimates with IPTW.	45
A-3	Relative bias of ATE estimates and the percent bias reduction of ATE estimates with TIPTW.	46
A-4	Relative bias of ATE estimates and the percent bias reduction of ATE estimates with OFPSM.	47
A-5	Relative bias of ATE estimates and the percent bias reduction of ATE estimates with PSS.	48
B-1	Relative bias of standard error estimates in the baseline.	49
B-2	Relative bias of standard error estimates with IPTW.	50
B-3	Relative bias of standard error estimates with TIPTW.	51
B-4	Relative bias of standard error estimates with PSS.	52
B-5	Relative bias of standard error estimates with OFPSM.	53
C-1	Empirical coverage rates of 95% confidence intervals across 1000 simulated data sets.	54
C-2	Proportion of the estimated ATE that is significant at α=.05 level across 1000 simulated datasets.	55

LIST OF FIGURES

Figure page

4-1 Relative bias of different standard error estimation methods in the baseline..... 36

4-2 Relative bias of different standard error estimation methods in the IPTW 37

4-3 Relative bias of different standard error estimation methods in the TIPTW..... 38

4-4 Relative bias of different standard error estimation methods in the PSS..... 39

4-5 Relative bias of different standard error estimation methods in the OFPSM 40

LIST OF ABBREVIATIONS

ATE	Average Treatment Effect
ATT	Average Treatment Effect on the treated
IPTW	Inverse Probability of Treatment Weighting
JK	Jackknife
NCES	National Center for Education Statistics
NSF	National Science Foundation
OFPSM	Optimal Full Propensity Score Matching
PS	Propensity Score
PSS	Propensity Score Stratification
SUTVA	Stable Unit Treatment Value Assumption
TIPTW	Truncated Inverse Probability of Treatment Weighting
TSL	Taylor Series Linearization
WLS	Weighted Least Squares Regression

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts in Education

THE PERFORMANCE OF PROPENSITY SCORE METHODS TO ESTIMATE THE AVERAGE TREATMENT EFFECT IN OBSERVATIONAL STUDIES WITH SELECTION BIAS: A MONTE CARLO SIMULATION STUDY

By

Sungur Gurel

August 2012

Chair: Walter Leite
Major: Research and Evaluation Methodology

We investigated the performance of four different propensity score (PS) methods to reduce selection bias in estimates of the average treatment effect (ATE) in observational studies: inverse probability of treatment weighting (IPTW), truncated inverse probability of treatment weighting (TIPTW), optimal full propensity score matching (OFPSM), and propensity score stratification (PSS). We compared these methods in combination with three methods of standard error estimation: weighted least squares regression (WLS), Taylor series linearization (TSL), and jackknife (JK). We conducted a Monte Carlo Simulation study manipulating the number of subjects and the ratio of treated to total sample size. The results indicated that IPTW and OFPSM methods removed almost all of the bias while TIPTW and PSS removed about 90% of the bias. Some of TSL and JK standard errors were acceptable, some marginally overestimated, and some moderately overestimated. For the lower ratio of treated on sample sizes, all of the WLS standard errors were strongly underestimated, as designs get balanced, the underestimation gets less serious. Especially for the OFPSM, all of

the TSL and JK standard errors were overestimated and WLS standard errors under estimated under all simulated conditions.

CHAPTER 1
INTRODUCTION

Estimating the effects of educational interventions using secondary data has become common in educational research because of the availability of various nationally-representative databases collected by agencies such as the National Center for Education Statistics (NCES) and the National Science Foundation (NSF) (Strayhorn, 2009). However, because assignment of participants to interventions in these national studies is not random, estimates of the effects of interventions are vulnerable to selection bias due to both observed and unobserved covariates (Shadish, Cook, & Campbell, 2002). In the last three decades, several methods emerged for estimating treatment effects and dealing with selection bias in studies that lack random assignment to treatment conditions (Heckman, 1978; Rosenbaum & Rubin, 1983; Abadie & Imbens, 2006; Heckman et al., 1997), which are referred collectively as observational studies. Propensity score methods are among the most commonly used methods in social science research in the analysis of observational studies. In order to reduce selection bias in treatment effect estimates, propensity score methods attempt to balance differences between treated and untreated participants on observed covariates. Rosenbaum and Rubin (1983) used the term propensity score (PS) for the first time and defined it as the predicted probability of treatment assignment given observed covariates. They found that if selection into treatment depends on observed covariates, the observed difference in treatment and control conditions at a propensity score level is an unbiased estimate of the average treatment effect (ATE) on that level (Rosenbaum & Rubin, 1983). Propensity scores can be used to reduce selection bias in the ATE by matching observations based on their similarity in PS, weighting observations with the

inverse of the PS, stratifying observations into homogenous groups based on PS (Stuart, 2010).

In addition to dealing with the problem of obtaining unbiased treatment effect estimates despite the existence of selection bias, educational researchers using large scale surveys also have to pay special attention to the estimation of standard errors, because changes in the sampling variability of the data due to the use of PS methods such as propensity score matching and stratification may require special methods to estimate standard errors such as bootstrapping, jackknife and Taylor-Series linearization (Stapleton, 2008).

Although several types of treatment effects have been defined in the literature (Guo & Fraser, 2010), the estimates most commonly found in the social sciences literature are the average treatment effect (ATE) and the average treatment effect on the treated (ATT) (Thoemmes & Kim, 2011). Although all PS methods can be used to estimate these two treatment effects, the specific implementation of a propensity score method differs depending on whether the ATE or ATT are of interest. There have been several studies comparing implementations of PS methods to estimate the ATT (Gu & Rosenbaum, 1993; Cepeda, Boston, Farrar, & Storm, 2003; Austin, 2010b; Harder, Stuart, & Anthony, 2010), but there has not been a study comparing all major propensity score methods for the estimation of the ATE. Therefore, the first objective of this study is to compare full optimal propensity score matching, propensity score stratification and inverse probability weighting with respect to their ability to reduce selection bias in estimates of the ATE. Because most studies comparing PS methods focused on treatment effect estimates and did not address estimation of standard errors, the

second objective of this study is to compare strategies for estimating standard errors for the ATE estimates obtained with each propensity score method.

CHAPTER 2
THEORETICAL FRAMEWORK

The Potential Outcomes Framework

Rubin's potential outcomes framework (1974) is commonly used to understand selection bias in observational studies. Its basic principle is that treated and control individuals have potential outcomes in both presence and absence of treatment. For instance, let the observed outcome of a treated participant *i* be Y_{iT}^{t} , while Y_{iT}^{c} is the potential outcome if this participant had been placed in the control group. Similarly, Y_{iC}^{c} is the observed outcome for the control group participant *i* and Y_{iC}^{t} is the potential outcome if this participant had been placed in the treatment group. In other words, a control group participant has a potential outcome under the treatment condition. Conversely, a treatment group participant has a potential outcome under the control condition. In the randomized studies, considering treatment assignment is random, the expected value of the potential outcomes of the treatment group $E[Y_{iT}^{c}]$ is equal to the expected value of observed outcomes of the control group $E[Y_{iC}^{c}]$. Similarly, the expected value of the potential outcomes of the control group $E[Y_{iC}^{t}]$ is equal to the expected value of observed outcomes of the treatment group $E[Y_{iT}^{t}]$. Therefore, the unbiased estimate of the treatment effect is $E[Y_{iT}^{t}]$ - $E[Y_{iC}^{c}]$.

In observational studies, due to non-randomness of the treatment assignment, groups may not be equivalent in the absence of treatment. Consequently, we cannot assume the $E[Y_{iT}^{c}] = E[Y_{iC}^{c}]$ and $E[Y_{iC}^{t}] = E[Y_{iT}^{t}]$. As stated earlier, many different treatment effects are defined within the potential outcomes framework but we will give special attention for those which are commonly used in the social science literature.

Average treatment effect on the treated is $ATT = E[Y_{iT}^{t}] - E[Y_{iT}^{c}]$, which is defined as the difference between the observed outcome of the participant under the treatment condition and the potential outcome under the control condition. In contrast, average treatment effect $ATE = E[Y_{i}^{t}] - E[Y_{i}^{c}]$, which is defined as the difference between the potential outcome for all individuals if they are exposed to the treatment condition and the potential outcome for all individuals if they are exposed to the untreated condition (Winship & Morgan, 1999). In other words, ATT is the effect for those in the treatment group; while ATE is the effect on both treated and untreated individuals (Stuart, 2010).

The estimate of the ATE based on the difference between the observed outcomes of the treated and control individuals will only be unbiased if the assignment to the treatment is independent of the potential outcomes. More formally, $T \perp \left\{Y_i^t, Y_i^c\right\} \mid X$ where Y_i^t is the potential outcome if treated, Y_i^c is the potential outcome if untreated, *X* is the all potential confounders, and *T* is the treatment assignment. This condition is known as strong ignorability of treatment assignment (Rubin, 1974). It is also necessary that the stable unit treatment value assumption (SUTVA) is met, which requires that the potential outcome of one unit is not affected by the particular treatment assignment of other units (Rubin, 2007). Random assignment meets these assumptions, but frequently in social research the randomized experiment is either unfeasible or unethical. In observational studies, both strong ignorability of treatment assignment and SUTVA may be violated, which leads to biased estimates of the ATE and poor internal validity of the study (Shadish, 2002). PS methods attempt to achieve strong ignorability of treatment assignment, under the assumption that SUTVA holds, by balancing the distributions of observed covariates between treatment and control

groups. Violations of SUTVA, such as when the decision of a parent to enroll a student in an educational intervention is affected by the enrollment status of the student's classmates, require special considerations in the estimation of propensity scores and implementation of the PS method that are discussed elsewhere (Arpino & Mealli, 2011; Thoemmes & West, 2011).

Propensity Score Methods for Reducing Selection Bias in ATE Estimates

The use of any PS method requires a multiple step process that starts with the selection of observed covariates that are related to selection into treatment conditions. The second step is to estimate propensity scores, which is most commonly accomplished with logistic regression, but other methods such as boosted regression trees (McCaffrey, Ridgeway, & Morral, 2004) can be used. The third step is to evaluate the common support area of the estimated propensity scores, which is the area of the propensity score distribution where values exist for both treatment and control groups (Guo & Fraser, 2010). Lack of common support for a certain area of the propensity score distribution restricts the generalizability of the estimates only to the sub-population for which common support exists. The fourth step is to verify the balance of the distribution of covariates given the propensity score method of choice. The fifth step is to apply the PS method in conjunction with a statistical method (e.g. ordinary least squares regression) to estimate the ATE and its standard error, and reach conclusions about statistical significant of the ATE. The last step is to evaluate the sensitivity of the results to possible omission of important covariates (Rosenbaum, 2010).

Inverse Probability of Treatment Weighting (IPTW)

Inverse probability weighting was introduced around the middle of 20th century by Horwitz and Thompson (1952) to account for the effect of the sampling design in survey

estimates. Robins, Hernan, and Brumback (2000) extended this concept to inverse probability of treatment weighting, to control for selection bias in observational studies. The simple idea behind IPTW is to weight subjects by the inverse of the conditional probability of being in the group that they are actually in. Formally, let T_i be the treatment indicator, with $T_i = 1$ indicating a member of the treatment group and $T_i = 0$ indicating a member of the control group. $\hat{e}_i$ is the estimated propensity score. To estimate the ATE, all individuals in the sample are given weights. For individual *i* the weight w_i is (Stuart, 2010):

$$w_i = \frac{T_i}{\hat{e}_i} + \frac{1 - T_i}{1 - \hat{e}_i} \quad \text{(2-1)}$$

IPTW models create a pseudo population where observations are replicated based on the weights, so that participants not only account for themselves, but also for those who have similar characteristics in the other group (Hernan, Hernandez-Diaz, & Robins, 2004). Neugebauer and van der Laan (2005) claimed that the performance of IPTW depends on the experimental treatment assignment assumption, which requires all of the weights are different from zero. They also found that if any treatment probability is close to zero, the new weighted sample may not be representative of the target population.

Truncated Inverse Probability of Treatment Weighting (TIPTW)

The IPTW method has been criticized regarding its performance when the weights are extreme or the propensity scores are extreme (Freedman & Berk, 2008). The extreme weights create overly influential observations and inflate the sampling variability of estimates. Several researchers came up with different solutions to solve this problem. Bemboom and van der Laan (2008) developed a data-adaptive selection

of truncation level for IPTW estimators. They were able to gain up to 7% efficiency in mean square error of estimates. Freedman and Berk (2008) replaced the weights that are greater than 20 with 20 and trimmed observations greater than 20. However, they concluded neither method was able to reduce the selection bias. Strumer, Rothman, Avorn, and Glynn (2010) found that trimming up to propensity scores that are more extreme than 2.5th and 97.5th percentiles reduces selection bias compared to not trimming any observations and trimming more observations.

Propensity Score Stratification (PSS)

PSS consists of creating strata containing individuals that are similar with respect to propensity scores (Stuart, 2010), where each strata should contain at least one treated and one untreated individual. PSS is usually accomplished by dividing the distribution of propensity scores into intervals of equal size. Stratification based on covariates to reduce selection bias was proposed by Cochran (1968), but Rosenbaum and Rubin (1984) showed that stratification into five strata based on the propensity scores reduces about 90% of the selection bias. In applied social science research, Thoemmens and Kim (2011) found that most studies use between 5 and 10 strata. Obtaining strata to estimate the ATE requires that all members of the sample are placed into a stratum, while stratum containing only untreated observations may be dropped in the estimation of ATT. Furthermore, estimating ATE requires the cases are weighted by the number of the individuals in each stratum based on the following formula:

$$w_{jt} = \frac{1}{2}\frac{N}{J * n_{jt}} \tag{2-2}$$

where j indexes strata, t indexes the treatment condition, N is the total sample size, J is the total number of strata, and n_{jt} is the total number of treated or untreated participants

in the stratum j. In contrast, ATT weights are created according to the number of treated individuals in each stratum.

Optimal Full Propensity Score Matching (OFPSM)

Full matching is a method of stratification in which the number of subclasses is shaped based on the data. When picking a treated unit randomly and from the same subclass of picking a control unit randomly, the expected difference between those two with respect to a certain measure of distance is Δ. All of the observations can be matched using a greedy algorithm, where observations are reviewed one-by-one to select the closest observations with respect to Δ without considering the minimization of the overall distance for the whole sample. The optimal full matching algorithm was established based on network flow theory to minimize Δ within matched sets by finding a minimum cost flow for the whole sample (Rosenbaum, 1989:1991). OFPSM is a special case of optimal full matching where the PS is used as the measure of distance in the matching procedure. Formally, let $\hat{e}_{it}$ be the fitted propensity score of the individual *i* exposed to treatment *t* and $\hat{e}_{jc}$ be the fitted propensity score of an individual *j* exposed to the control condition *c*, given observed covariates. $\Delta_{ij} = \left|\hat{e}_{it} - \hat{e}_{jc}\right|$ is the distance between propensity scores. Rosenbaum (1991) found that there is always a full matching that is optimal so that $\sum_{i,j}^{T,C} \Delta_{ij}$ is minimized where *T* is number of treated participants and *C* is number of control participants in a sample. Estimating the ATE with OFPSM requires weights calculated in the same way as in PSS (see Equation 2-2).

Standard Error Estimation with Propensity Score Methods

Weighted least squares regression (WLS) can be used to obtain ATE estimates and standard errors with all PS methods presented above (Schafer & Kang, 2008). This method can be implemented through applying proper weights for each PS method while fitting the regression models. For the IPTW and TIPTW methods, weights are obtained using Equation 2-1. For the OFPSM and PSS methods, weights are shaped based on the number of strata and the number of treated and untreated individuals according to Equation 2-2. Once weights are shaped, standard errors in the WLS are estimated with the following formula;

$$SE(\hat{B}_1) = \sqrt{\frac{\sum_{i=1}^{n}\left(E_i / w_i\right)^2}{n}} \qquad (2\text{-}3)$$

where E_i is the residual and w_i is the weight for the observation *i* and *n* is the sample size (Fox, 2008).

Several other methods can be used to obtain standard errors of ATE estimates from PS methods, such as Taylor Series Linearization, Jackknife, and Bootstrapping (Rodgers, 1999). These methods have not been researched extensively with propensity score methods. Taylor series linearization (TSL) can be used to obtain the variance of a statistic via approximating the estimator by a linear function of observations (Wolter, 2007). Formally, let $U_i(ATE)$ be a function of the data for observation i and ATE, and the true population ATE^* solves the following equation;

$$\sum_{i=1}^{N} U_i(ATE^*) = 0 \qquad (2\text{-}4)$$

then, in a complex sample we are able to define $\hat{ATE}$ as solving the weighted sample equation

$$\sum_{i=1}^{N} \hat{U}_i(ATE) = 0 \tag{2-5}$$

Variance of the ATE is defined as follows applying delta method (Binder, 1983);

$$\hat{\text{var}}[\hat{ATE}] \approx \left(\sum_{i=1}^{n} \frac{\partial U_i(\hat{ATE})}{\partial ATE} \right)^{-1} \text{cov}\left[\sum_{i=1}^{n} \hat{U}_i(\hat{ATE}) \right] \left(\sum_{i=1}^{n} \frac{\partial U_i(\hat{ATE})}{\partial ATE} \right)^{-1} \tag{2-6}$$

For the observation *i* let $\hat{B}_0$ be the intercept, $\hat{B}_1$ be the slope of the regression equation where $\hat{B}_1$ is estimated ATE, w_i be the weight, and $\hat{\bar{x}}$ be the mean of x. Using Taylor series linearization, standard error of ATE is defined as (Lohr, 1999);

$$SE(\hat{B}_1) = \sqrt{\frac{\hat{V}(\sum_{i=1}^{n} w_i(y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \hat{\bar{x}})}{\left[\sum_{i=1}^{n} w_i x_i^2 - \frac{\left(\sum_{i=1}^{n} w_i x_i \right)^2}{\sum_{i=1}^{n} w_i} \right]}} \tag{2-7}$$

Jackknife and bootstrapping are both based on resampling from the original data. The most common implementation of the jackknife (JK) is the delete-1 jackknife, where at each iteration, one member of the sample is removed randomly and the parameters of interest are estimated using replicated weights, which are recalculated after removing the observation. For delete-1 jackknife let the w_i be the initial weight for an observation *i* and *n* be the sample size. Depending on the PS method selected, w_i may be IPTW weight obtained from Equation.2-1 or truncated weights;

$$w_{ik} = \begin{cases} 0 & \text{if the observation unit i is deleted at iteration k} \\ \dfrac{n}{n-1} w_i & \text{if the observation unit i is not deleted at iteration k} \end{cases} \tag{2-8}$$

At each iteration, the ATE ($\hat{B}_{1k}$) which is the parameter of interest is re calculated. The standard error will be as follows (Lohr, 1999);

$$SE(\hat{B}_1) = \sqrt{\frac{n-1}{n}\sum_{k=1}^{k}(\hat{B}_{1k} - \hat{B}_1)^2} \quad (2\text{-}9)$$

Delete-n jackknife is used when a whole stratum is deleted. Let w_{jt} be the weight for a particular stratum weight that are calculated using Equation.2-2, *n* be the sample size, and n_{jt} be the strata size for strata *j*. The same rule applies for the estimation of standard errors using delete-n jackknife but just the jackknife weights are estimated by the following way;

$$w_{jtk} = \begin{cases} 0 & \text{if all of the observation units in stratum jt are deleted at iteration k} \\ \dfrac{n}{n-n_{jt}} w_{jt} & \text{if any of the observation units in stratum jt are not deleted at iteration k} \end{cases} \quad (2\text{-}10)$$

At each iteration, the ATE ($\hat{B}_{1k}$) which is the parameter of interest is re calculated using new weights. The standard error is obtained using the following equation (Lohr, 1999);

$$SE(\hat{B}_1) = \sqrt{\frac{1}{k-1}\frac{n-n_{jt}}{n}\sum_{k=1}^{k}(\hat{B}_{1k} - \hat{B}_1)^2} \quad (2\text{-}11)$$

Bootstrapping consists of resampling *k* times with replacement from the original sample to create samples of the same size as the original sample, and estimating parameters with the *k* resampled datasets. Weights are adjusted for each bootstrapped sample by multiplying how many times a particular observation is selected to be in the resample. Standard errors are simply standard deviation of the parameter estimates across the multiple samples. Formally (Lohr, 1999);

$$SE(\hat{B}_1) = \sqrt{\frac{1}{k-1}\sum_{k=1}^{k}(\hat{B}_{1k} - \hat{B}_1)^2} \quad (2\text{-}12)$$

However, Abadie and Imbens (2008) demonstrated a mathematical proof and ran a simulation about the variance estimation accuracy of ATT estimates at nearest neighbor matched samples. They argued bootstrapping is not appropriate for matched

data because when the ratio of treated on the sample size is greater than about .42, standard bootstrap does not provide a valid estimate of the asymptotic variance of the ATT and increasing sample size is not a solution either. However, Abadie and Imbens claimed that bootstrapping provides valid inference at propensity score weighting. Due to the fact that the propensity score weighting estimator is asymptotically linear, bootstrapping was not included in this study because it does not apply all the conditions that we have investigated. Abadie and Imbens didn't extend their study to the accuracy of standard errors that are estimated through the jackknife method. Because we don't have enough evidence on whether the same problem applies to the jackknife method or not we did not exclude the jackknife method from the study.

Comparison of Propensity Score Methods' Performances

Much research has been conducted comparing different structures, distances and algorithms for propensity score matching or probability weighting to estimate the treatment effects (Gu & Rosenbaum, 1993; Cepeda, Boston, Farrar, & Storm, 2003; Lunceford & Davidian, 2004; Austin, 2009a; Austin, 2009b; Austin, 2010a; Harder, Stuart, & Anthony, 2010). For this study, it is particularly relevant that both Gu and Rosenbaum (1993) and Cepeda et al. (2003) found the optimal matching consistently outperforms matching with a greedy algorithm, which is the most commonly used algorithm for PS matching. For this reason, we did not investigate the greedy algorithm further. Austin (2009a) found that matching on the propensity score within a specified caliper and IPTW methods removes more systematic differences between groups than PSS and covariate adjustment. Austin (2010a) also found that a doubly robust IPTW method (i.e., where the IPTW is used as both a weight and a covariate) works better than PSS, matching on propensity score, IPTW, and covariate adjustment methods in

terms of bias, variance estimation, coverage of confidence intervals, mean squared error, and Type I error rates. However, Austin only compared these methods under the condition of a binary outcome. Also, Austin (2009b) evaluated standard error estimation methods for propensity score matching and found that methods that considered the matched nature of the data resulted in smaller bias of standard errors and actual Type I error rates closer to the nominal Type I error rate. Harder et al. (2010) compared the interaction of three PS estimation models (e.g., multivariable logistic regression, multivariable logistic regression with product terms, and nonparametric generalized boosted modeling) and five PS methods (e.g.,1:1 greedy matching, full matching, weighting by odds, stratification, and IPTW). Because in their study the ATT and ATE estimates were similar, they made comparisons across methods that estimate the ATT and the ATE. Their results indicated that using nonparametric generalized boosted modeling to estimate propensity scores and 1:1 greedy matching provides better covariate balance for the majority of the covariates. Lunceford and Davidian (2004) compared IPTW with PSS, with respect to bias removal in the ATE. Both theoretical and empirical results indicated that using a fixed number of strata lead to biased estimates of the ATE.

In this study, we address the scarcity of information in the literature about the relative performance of PS methods for estimating the ATE and its standard error with the following research questions:

1. Which propensity score method (OPSM, IPTW, TIPTW, and PSS) performs best with respect to unbiased estimation of ATE under conditions with different sample size, and ratio of treated to total sample size?
2. Which method (WLS, TSL, and JK) produces the most accurate standard errors when combined with different propensity score methods (OPSM, IPTW, TIPTW, and PSS)?

3. Which propensity score method (OPSM, IPTW, TIPTW, and PSS) leads to the most power to test the ATE?

CHAPTER 3
METHOD

In order to answer the research questions, a Monte Carlo simulation study was conducted using the R.2.14.0 program (R Development Core Team, 2011).

Data Simulation

The data was simulated based on manipulating sample size and the proportion of treated individuals. Monte Carlo simulation studies to compare PS methods by Gu and Rosenbaum (1993), Freedman and Berk (2008), and Austin (2009a) used 1,000 as the only sample size. By manipulating sample size, we were able to determine whether there were differences between the PS methods in terms of power to test the ATE. We simulated data with sample sizes equal to 500, 1000, and 2000.

We generated data where the proportion of the sample that was treated was set at 1/10, 1/7, 1/4, 1/3, and 1/2. These conditions are an extension of Gu and Rosenbaum's (1993) study, which only examined ratios of 1/7, 1/4, 1/3.

To measure the common support area of the propensity score, we used an overlap measure that is similar to the Cohen's (1988) U_1 function, which is proportion of non-overlap of the distributions. Let A and C be the area of non-overlap and B be the overlap area of the logit of propensity scores. Then,

$$U_1 = \frac{A + C}{A + B + C} \qquad (3\text{-}1)$$

where increases in U1 correspond to decreases in the area of common support. In the simulated conditions, the mean U1 ranged from .131 to .280. As the sample size increased or the ratio of treated to sample size increased, overlap also improved.

In order to obtain reasonable population parameters to simulate data, we took estimates from the 2007-2008 School Survey on Crime and Safety (SSOCS) survey results. SSOCS is a nationally representative survey conducted by the United States Department of Commerce (NCES, 2011).The covariates were number of students transferred from school, typical number of classroom changes, percentage of students below 15th percentile standardized tests, and the total number of transfers to specialized schools. The grouping variable was whether outside school diciplinery plan available or not and the outcome was total number of students involved in specified offenses.

We generated multivariate-normally distributed covariates for the simulation study using the Mass 7.3-16 package in R (Venables, & Ripley 2002). The first step of data simulation was to simulate the covariates $X_{1i}, X_{2i}, X_{3i}, X_{4i}$, which where normally distributed with population means of zero and population covariance matrix equal to

$$\begin{bmatrix} 1.00 & .145 & -.004 & .125 \\ .145 & 1.00 & .001 & .467 \\ -.004 & .001 & 1.00 & .061 \\ .125 & .467 & .061 & 1.00 \end{bmatrix}$$

At the second step, we simulated residuals of the outcome regression. Residuals were simulated from a normal distribution with mean of zero and standard deviation of 166.208. The population standard deviation of the residuals was defined so that the population R^2 for the outcome regression was .211. Once the covariates and the residual of the outcome were simulated, we obtained the potential control outcomes Y_C and potential treatment outcomes Y_T for all individuals in the sample based on following equations:

$$Y_{Ci} = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i$$
$$Y_{Ti} = Y_{Ci} + ATE \quad \text{(3-2)}$$

The population values of the coefficients β_0, β_1, β_2, β_3, and β_4 were *0, 16.221, 58.642, 15.704,* and *33.601.* The population value of the ATE was 20, which corresponds to a Cohen's effect size of .085 indicating that this is a small effect.

The next step was to determine which individuals in the simulated samples were exposed to treatment. The population model for treatment assignment was:

$$Log(\Pr(T_i = 1 \mid X_{1i}, X_{2i}, X_{3i}, X_{4i})) = \log(rt/(1-rt)) + \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{4i}) \quad \text{(3-3)}$$

where the population values of $\gamma_0, \gamma_1, \gamma_2, \gamma_3,$ and γ_4 were 0, .127, .137, .166, and .101, and *rt* is the ratio of treated to the sample size. The strength of the selection bias was defined based on the McKelvey & Zavoina pseudo R^2 , and its population value for the simulated data was .028 (McKelvey, & Zavoina 975). The population model for treatment assignment defines the effect of covariates on the logit of probability of being in the treatment group. We included the odds of being in the treated group $rt/(1-rt)$ in the treatment assignment model so that we could control ratio of treated to the sample size. Finally, we compared the probability of being in the treatment group given observed covariates to random number that was obtained from a uniformly distributed population with a maximum value of one and a minimum value of zero for each observation. A case in each simulated sample was defined as a treated if the probability of being in the treatment group was greater or equal to the random number. Otherwise, the case was defined as untreated.

We did not manipulate the number of covariates because Gu and Rosenbaum (1993) concluded that as long as treatment assignment mechanism is modeled

completely, the number of covariates does not affect the performance of the propensity score estimation process, with the exception of potential problems of multicollinearity of covariates and convergence problems. We used four continuous covariates that were both related to the outcome and treatment assignment in this simulation study. Since we simulated data based on four covariates and estimated propensity scores using all four covariates, the assumption that the treatment assignment is modeled completely was met for all conditions in the study.

Estimation of ATE and Standard Errors

We simulated 1,000 datasets per condition. We analyzed the simulated data according to the following steps for each dataset:

1. We estimated the PS for each individual using logistic regression.
2. We estimated Cohen's U_1 value to measure the area of common support.
3. We estimated ATE and standard error ignoring selection bias to represent the baseline with the following model:

$$Y_i = \pi_o + \pi_1 T_i \tag{3-4}$$

   where π_o is an intercept, T_i is a treatment indicator, and π_1 is the ATE estimate.

4. Using Equation 3-4, we estimated ATE and standard errors using weights equal to 1. We re-estimated standard errors using the TSL method using equation 2-7.We implemented delete-1 JK method the survey 3.26.1 (Lumley, 2011) library in R. Weights are re estimated using Equation 2-8 and standard errors are re-estimated using Equation 2-9.
5. Using Equation 3-4, we estimated ATE and standard errors using IPTW weights with weighted least squares (WLS) estimation in the Equation 2-3. The weights used in the analysis are formed with the Equation 2-1 that was provided earlier. We re-estimated standard errors using the TSL method using equation 2-7 We implemented delete-1 JK method. Weights are re estimated using Equation 2-8 and standard errors are re-estimated using Equation 2-9.
6. We replaced the weights which were greater than 99th percentile of the IPTW with the 99th percentile and created TIPTW weights. We estimated ATE and

standard error with Equation 3-4 using TIPTW weights with WLS estimation in the Equation 2-3. We re-estimated standard errors using the TSL method using equation 2-7 We implemented delete-1 JK method. Weights are re estimated using Equation 2-8 and standard errors are re-estimated using Equation 2-9.

7. We grouped the treated and control individuals into five strata, based on similarity in PSs using the MatchIt 2.4-20 (Ho, Imai, King, Stuart, 2007) library in R. We decided to create 5 strata because that is most commonly-used number of strata by the social science researchers. (Thoemmes & Kim, 2011). We estimated the ATE and standard errors using weighted least squares estimation in the equation 2-3, with weights are calculated based on Equation 2-2 and used the model in the following equation;

$$Y_i = \pi_o + \pi_1 T_i + \sum_{k=1}^{K-1} \lambda_k S_{ki} \qquad \text{(3-5):}$$

where S_{ki} is a dummy-coded indicator of membership in stratum *S*. We re-estimated the standard errors using the TSL method based on the Equation 2-7. We implemented delete-n JK method in the survey 3.26.1 (Lumley, 2011) library in R where one stratum is deleted at each iteration. Weights are re estimated using Equation 2-10 and standard errors are re-estimated using Equation 2-11. While analyzing the data we assumed data is stratified in nature.

8. We grouped the treated and control individuals based on similarity in PSs into a data-defined number of strata using the OFPSM algorithm implemented in optmach 0.7-1 (Hansen, & Fredrickson, 2009) library in R. We estimated ATE based on Equation 3-5 with the weights that are obtained using Equation 2-2.. Standard errors using WLS estimation based on the Equation 2-3. We implemented delete-n JK method where one stratum is deleted at each iteration. Weights are re estimated using Equation 2-10 and standard errors are re-estimated using Equation 2-11. While analyzing the data we assumed data is stratified in nature.

Analyses

We have compared the PS methods in terms of relative bias of ATE estimates and percent bias reduction of ATE estimates. We compared the standard error estimation methods in terms of relative bias of standard errors, and coverage of confidence intervals. For the methods resulting in acceptable bias of standard errors, we estimated the power to test the ATE. The relative bias of the ATE was calculated with $B(\hat{\theta}) = (\bar{\hat{\theta}} - \theta)/\theta$, where $\bar{\hat{\theta}}$ is the mean of the ATE estimates for all iterations of one

condition and θ is the population ATE. If the absolute value of the estimated $B(\hat{\theta})$ is larger .05, the bias is considered unacceptable (Hoogland & Boomsma, 1998). Despite the fact that this criterion applies within the Structural Equation Framework and we do not know whether this applies treatment effect estimation procedures or not, we used this criteria as a rule of thumb.

Because the magnitude of the relative bias of the ATE depends not only on the difference between the mean ATE and population ATE, but also the size of the ATE, we also evaluated the percent bias reduction which is defined as:

$$PBR(\hat{\theta}) = \frac{B(\hat{\theta})_{baseline} - B(\hat{\theta})_{method}}{B(\hat{\theta})_{baseline}} x100 \quad (3\text{-}6)$$

where $B(\hat{\theta})_{method}$ is the mean relative bias of using a particular method and $B(\hat{\theta})_{baseline}$ is the initial bias (Cochran & Rubin, 1973; Steiner, Cook, Shadish, & Clark, 2010).

The relative bias of the standard error is $B(S_{\hat{\theta}}) = [\overline{S}_{\hat{\theta}} - SD(\hat{\theta})] / SD(\hat{\theta})$, where $\overline{S}_{\hat{\theta}}$ is the mean of the estimated standard errors of ATE and $SD(\hat{\theta})$ is the empirical standard error, which is the standard deviation of estimated ATE. If the absolute value of the estimated $B(S_{\hat{\theta}})$ is larger than.1, the bias is considered unacceptable (Hoogland & Boomsma, 1998).

We estimated the power by calculating the proportion of ATE that are statistically significant at $\alpha = .05$ level for each condition. We also calculated the proportion coverage of confidence intervals which is the proportion of iterations where the population ATE falls within the 95% confidence interval for the estimated ATE (Austin, 2009b).

CHAPTER 4
RESULTS

Table A-1 shows that when we ignored the selection bias, the mean estimated ATE became more than twice the size of the population ATE. Tables A-2 through A-5 present the relative bias of the estimated ATE and percent bias reduction for each PS method. We found that with IPTW, the only unacceptable relative bias was when the sample size was 500 and the simulated ratio of treated over sample size was as low as .1. In terms of removing bias, in almost all conditions, IPTW removed more than 98% of the initial bias. On the other hand, TIPTW was not as effective as IPTW in terms of removing bias, especially when the ratio of treated to sample size was small. As the sample size increased, TIPTW removed more bias, but this difference was small. As the ratio increased, TIPTW removed more bias and when the ratio became .5, TIPTW worked almost as well as IPTW in terms of reducing the selection bias in the estimate of ATE. As it is shown in Table A-4, OFPSM worked as well as IPTW in all conditions. When the sample size was 500 and the ratio of treated to sample size was 1/10, the remaining bias was marginally unacceptable but as the ratio increased, all ATE estimates became unbiased. Finally, Table A-5 shows that the PSS method failed to reduce the selection bias to acceptable levels in all conditions. It is reasonable to say that, among the 4 methods that are investigated, IPTW and OFPSM removed almost all of the selection bias in the estimated ATE, while TIPTW provided acceptable results only with proportions of treated individuals approaching 50% and the PSS never provided adequate results. Overall, we found that the performances of the PS methods tended to improve as the sample size and the ratio of treated to sample size increased.

When we look at Tables from B-1 through B-5, we see that the sample size does not have a sizeable effect on the bias of the standard error estimates. Figures 4-1 through 4-5 are established by collapsing biases across the sample sizes. Table 4-1 shows that the relative biases of the standard error of the estimated ATEs in the baseline condition were acceptable in all conditions with TSL and JK methods. However, WLS standard error were biased when the simulated ratio of treated to sample size was lower than .25. For WLS, the underestimation of standard errors declined as the ratio of treated to sample size approached .5. Figure 4-2 depicts the bias of the standard errors of ATE estimates for the IPTW method with different approaches to estimate standard errors. We saw a similar pattern to the baseline model using WLS. For the two lowest simulated ratios of treated to sample size, standard errors were underestimated, but underestimation did not occur with larger ratios. In contrast, TSL estimated standard errors accurately for the lowest ratio, but for the larger ratios this methods overestimated standard errors either marginally or moderately. The jackknife provided marginally more extreme standard errors than TSL did. The same pattern was observed for the TIPTW method. The only difference is that TSL and JK produced more accurate estimates of standard errors. In fact, for the lowest two ratios, TSL and JK produced acceptable standard errors but for the other three ratios, they did not produce standard errors with acceptable levels of bias.

For the PSS, WLS produced again underestimated standard errors with the lowest two ratios and marginally acceptable underestimated standard errors with a ratio of .25. For the other ratios, WLS standard errors were acceptable. In contrast, TSL and JK produced overestimated standard errors for most of the conditions. However, for the

first smallest and largest ratios the biases were either within acceptable levels or marginally overestimated. For the other ratios of treated to total sample size , TSL and JK marginally or moderately overestimated the standard errors. For the OFPSM, standard error estimates obtained with WLS were underestimated in all conditions. Also, TSL and JK provided standard errors that were marginally or moderately overestimated, regardless of the sample size and the simulated ratio of treated to sample size.

Considering WLS provided substantially under estimated standard errors for the two smallest ratios of treated to sample size and TSL produced more accurate standard errors than JK overall, coverage of confidence intervals and power analysis is conducted using TSL standard errors. For each iteration, we estimated 95% confidence interval for each estimated ATE in the iteration. Results are presented in Table C-1. For all conditions, the population ATE was within that confidence interval for more than 95% of the time.

Based on Table C-2, PSS was the most powerful method to test the ATE for the most of the simulated conditions. The second most powerful PS method was TIPTW. IPTW was the third most powerful method, and finally OFPSM was the least powerful. As expected, the power increased as the sample size increased, but also increased as the simulated ratio of treated to sample size increased.

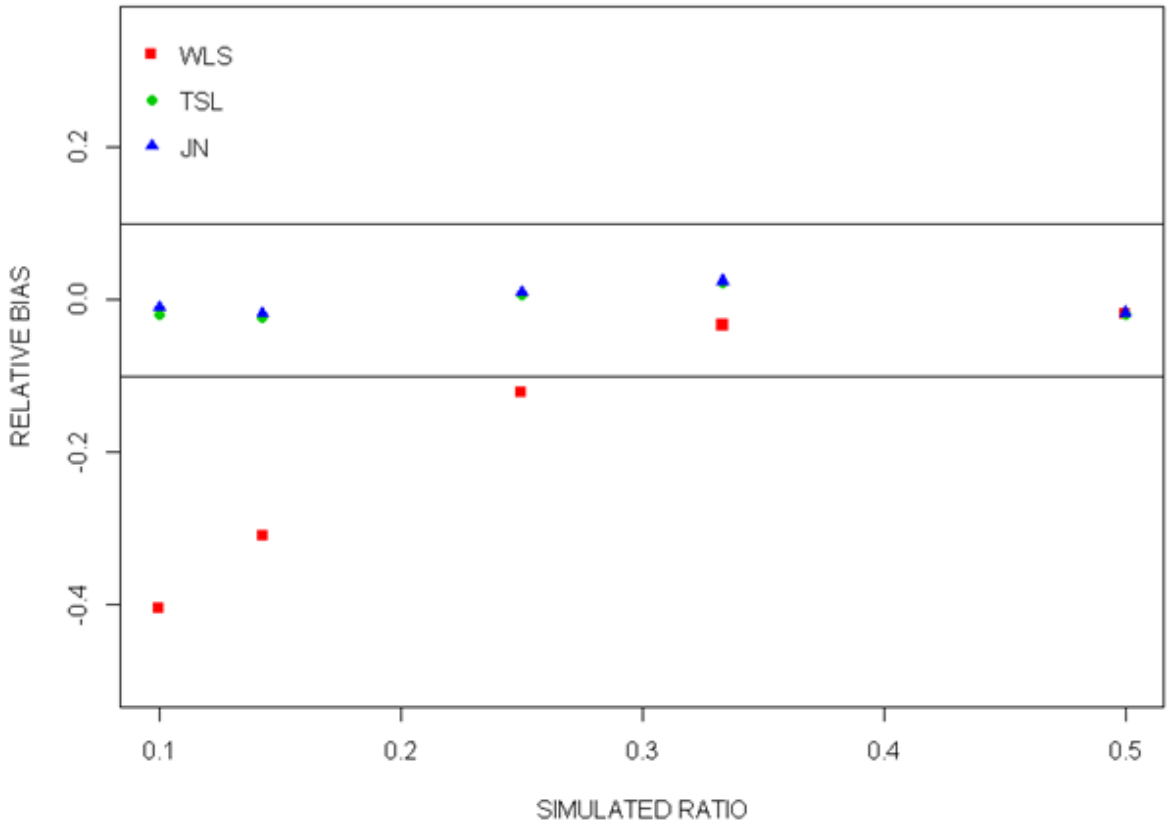

Figure 4-1. Relative bias of different standard error estimation methods in the baseline

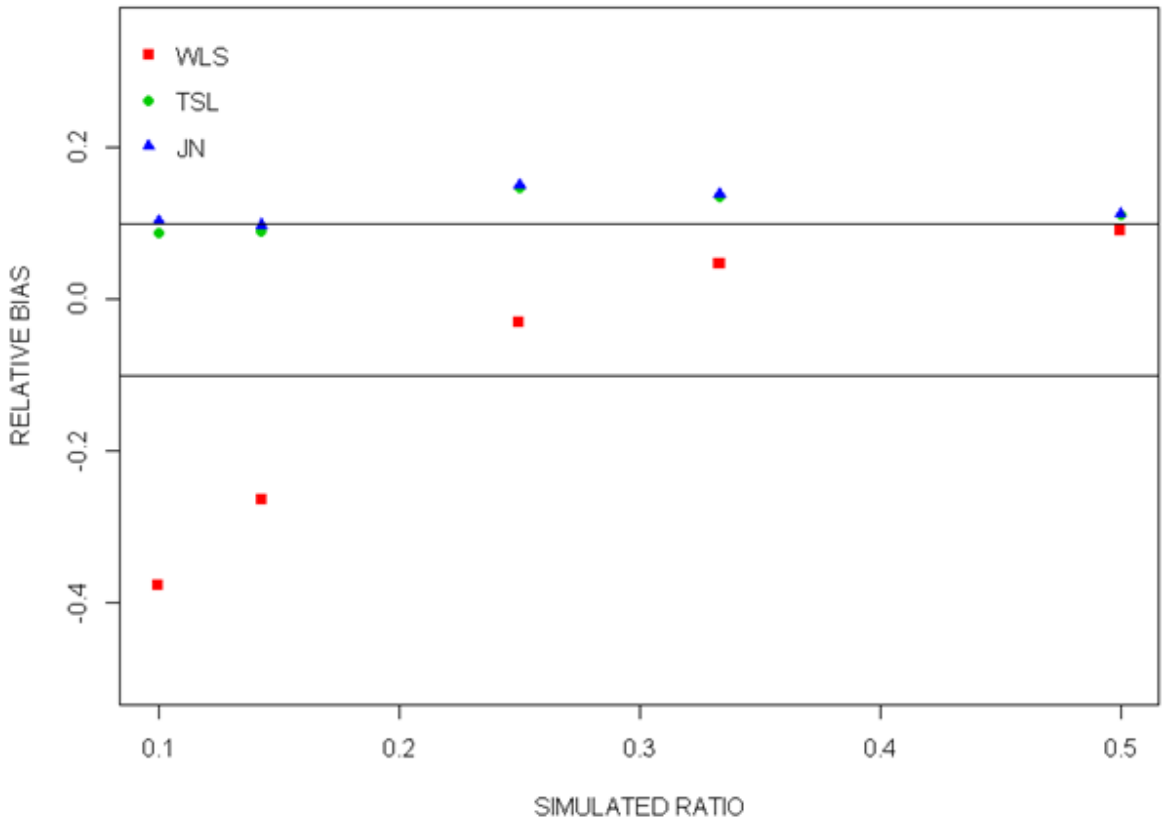

Figure 4-2. Relative bias of different standard error estimation methods in the IPTW

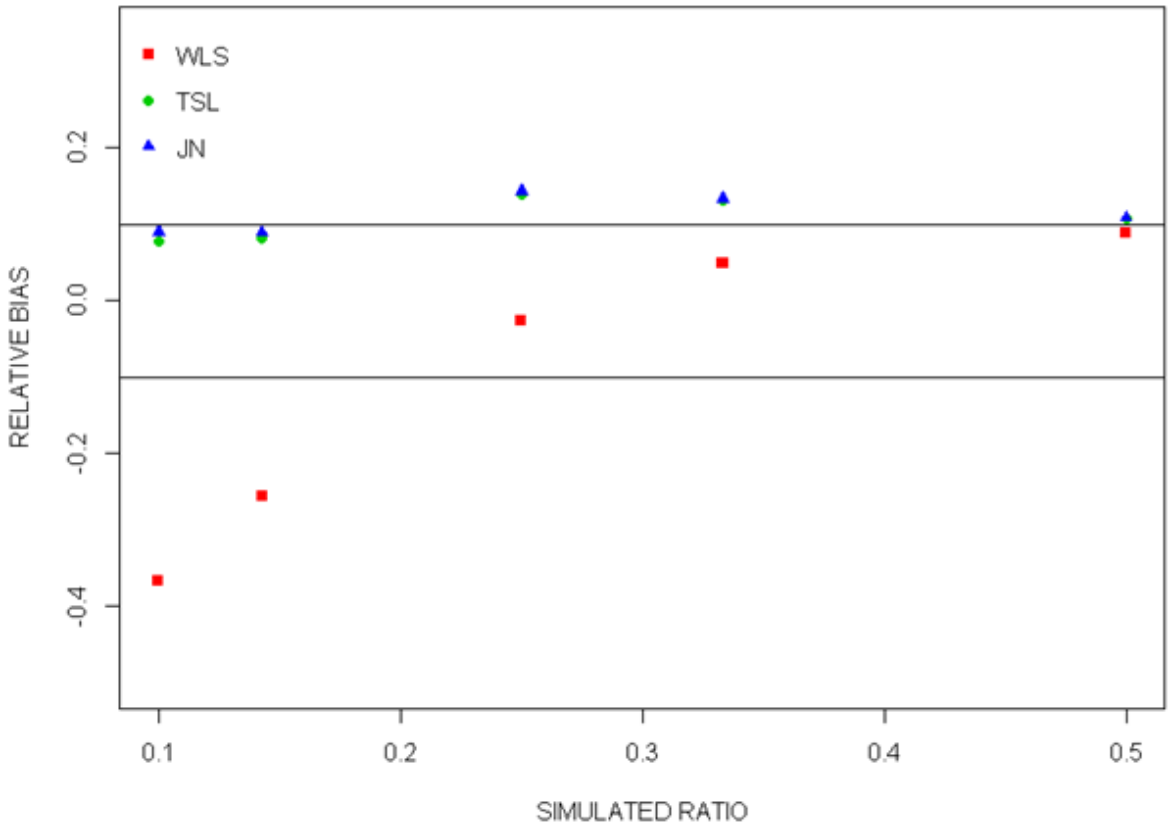

Figure 4-3. Relative bias of different standard error estimation methods in the TIPTW

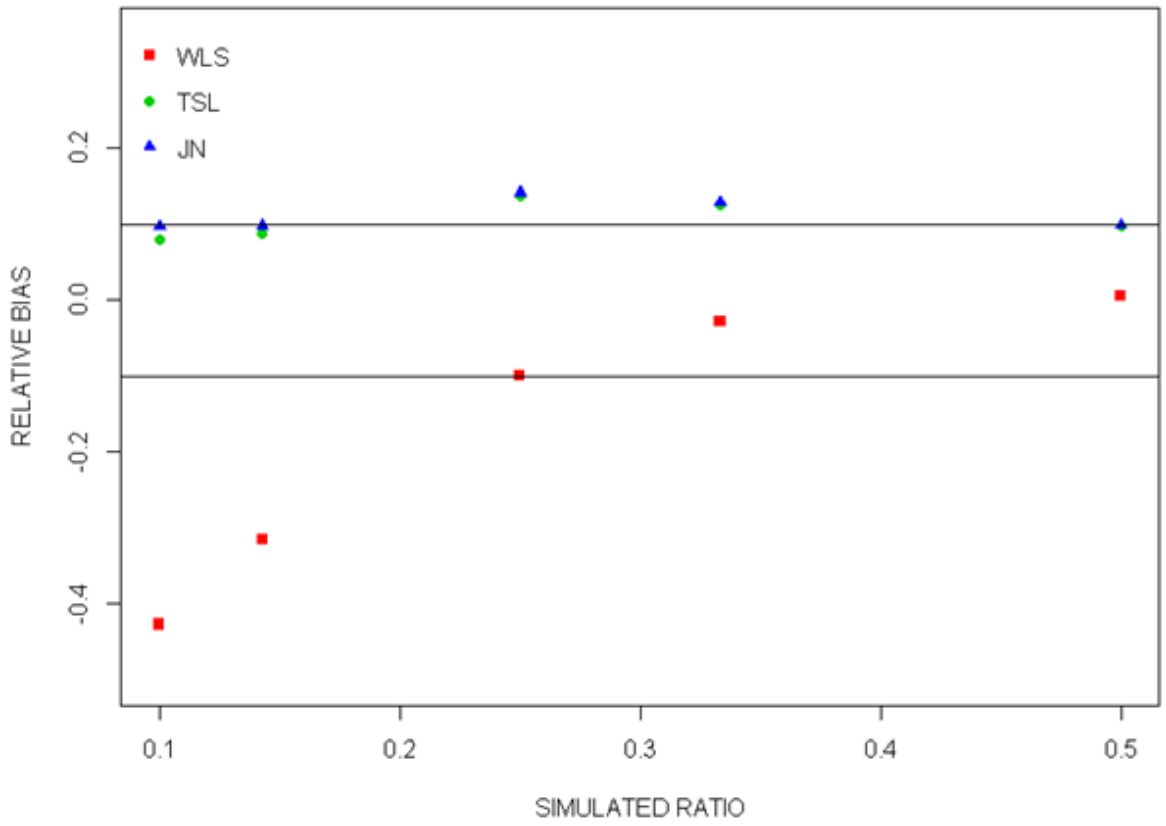

Figure 4-4. Relative bias of different standard error estimation methods in the PSS

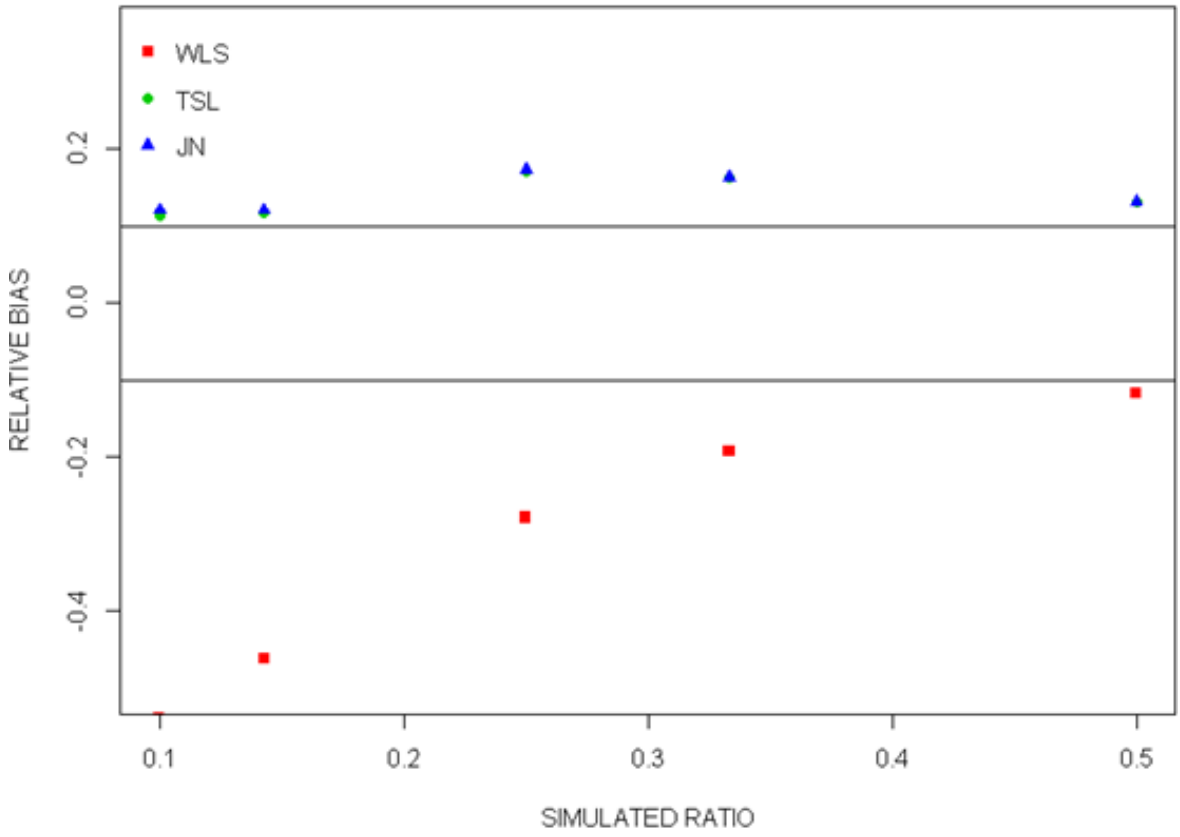

Figure 4-5. Relative bias of different standard error estimation methods in the OFPSM

CHAPTER 5
DISCUSSION

Our first research question asked whether a PS method outperformed the other in terms of removal of selection bias. With large levels of selection bias, IPTW and OFPSM removed almost all of the bias in every condition. OFPSM performed marginally better than IPTW overall. TIPTW was the third best method in terms of removing the selection bias and PSS was the worst method. Austin (2010a) found that IPTW removes more bias than PSS when the outcome is binominal. Cochran (1968) demonstrated that using 5 strata removes about 90% of the selection bias, which agrees with the percent bias reduction using PSS method that we found. We also found that as the ratio of treated to sample size increases from .1 to .5, the methods tend to remove more bias. This finding is partially consistent with Rosenbaum and Rubin's (1993) findings that as the number of possible matches increases, selection bias is reduced. Because we estimated the ATE, two-way matching is performed (i.e., treated to control and control to treated). In this case the number of available matches is maximum when the ratio of treated to sample size is .5. We also observed that as the sample size increases, all methods become more efficient in terms of bias reduction.

The second research question was about the accuracy of estimates of the standard errors. We are not able to conclude that a single standard error estimation procedure works best in every simulated condition. Data suggested that, TSL and JK tend to overestimate standard errors. Among these two methods, TSL is marginally more accurate than JK. Since TSL is a computation-based method, and JK is replication-based method, if TSL estimates are available, they should be more accurate than JK estimates. WLS tends to underestimate the standard errors. Even if the weights

are correct, WLS ignores the stratified nature of the data. This leads to underestimated standard errors when the strata are a key feature of the analysis such as in OFPSM and PSS. In addition, as the population ratio increases, estimates of standard error come within acceptable levels.

Our final research question was about the coverage of confidence intervals and power. For all simulated conditions, percent coverage of confidence intervals was greater than 95%. As expected, as the sample size increases, power increases. Also as expected, balanced designs are more powerful than unbalanced designs are (Hsieah, Blooch, & Larsen, 1998). Therefore, when the simulated ratio of treated to sample size is .5, we have more power to test the ATE.

In this study, we have investigated conditions where only the ignoring of treatment assignment was violated. Our results are not generalizable to conditions where SUTVA may be potentially violated. If this is the case, researchers are recommended to use appropriate model-based solutions such as multilevel modeling (Arpino & Mealli, 2011; Thoemmes & West, 2011). Because we estimated the ATE, the overlap between treated and the untreated individuals' PS distribution becomes important. Lechner (2008) concluded that poor overlap creates bias in the estimates of treatment effects. Ignoring the lack of common support may be misleading because groups may not be comparable. Dropping observations that are outside of the common support area in any group is also not desirable for the estimation of ATE because it conflicts with the definition of ATE. When there is a stratum with neither treated nor untreated participant, the weight for that stratum becomes zero and that stratum is ignored in the ATE estimation. This may happen while estimating ATE using OFPSM or PSS. In this

simulation, when one iteration dropped any observation due to the limited overlap using any of the methods, we ignored that iteration and re-simulated the data. For this reason, these results are not generalizable to the conditions where the overlap is poor and there is no available match for many of the observations.

One other limitation of the results is the weights that we used for the IPTW and TIPTW ranged from 1 to 20 for most of the conditions. We did not investigate conditions that have extreme weights such as 100 or more. The generalizability of our results is limited to the conditions that are not heavily affected by extreme weights. Also, additional simulation work should be performed to explain the reasons behind the standard errors not being accurate and possible solutions for this problem.

APPENDIX A
RELATIVE BIAS OF ATE ESTIMATES

Table A-1. Relative bias of ATE estimates in the baseline condition

Sample size	Simulated ratio of treated to total sample size	Relative bias of ATE estimates
500	1/10	1.195
	1/7	1.162
	1/4	1.181
	1/3	1.149
	1/2	1.155
1000	1/10	1.155
	1/7	1.152
	1/4	1.193
	1/3	1.153
	1/2	1.188
2000	1/10	1.219
	1/7	1.170
	1/4	1.168
	1/3	1.162
	1/2	1.119

Note. Biases that are greater than 0.05 or less than -0.05 are unacceptable.

Table A-2. Relative bias of ATE estimates and the percent bias reduction of ATE estimates with IPTW.

Sample size	Simulated ratio of treated to total sample size	Relative bias of ATE estimates	Percent Bias Reduction of the ATE estimates
500	1/10	0.073	93.930%
	1/7	0.008	99.321%
	1/4	0.039	96.716%
	1/3	0.001	99.932%
	1/2	-0.014	101.218%
1000	1/10	-0.020	101.736%
	1/7	-0.002	100.136%
	1/4	0.023	98.090%
	1/3	-0.009	100.821%
	1/2	0.021	98.262%
2000	1/10	0.022	98.207%
	1/7	0.000	99.986%
	1/4	0.004	99.693%
	1/3	0.012	98.971%
	1/2	-0.029	102.610%

Note. Biases that are greater than 0.05 or less than -0.05 are unacceptable.

Table A-3. Relative bias of ATE estimates and the percent bias reduction of ATE estimates with TIPTW.

Sample size	Simulated ratio of treated to total sample size	Relative bias of ATE estimates	Percent Bias Reduction of the ATE estimates
500	1/10	0.251	78.978%
	1/7	0.131	88.765%
	1/4	0.106	91.021%
	1/3	0.047	95.925%
	1/2	0.015	98.687%
1000	1/10	0.147	87.260%
	1/7	0.113	90.152%
	1/4	0.086	92.821%
	1/3	0.035	96.921%
	1/2	0.047	96.060%
2000	1/10	0.186	84.718%
	1/7	0.112	90.395%
	1/4	0.063	94.586%
	1/3	0.054	95.322%
	1/2	-0.005	100.416%

Note. Biases that are greater than 0.05 or less than -0.05 are unacceptable.

Table A-4. Relative bias of ATE estimates and the percent bias reduction of ATE estimates with OFPSM.

Sample size	Simulated ratio of treated to total sample size	Relative bias of ATE estimates	Percent Bias Reduction of the ATE estimates
500	1/10	0.071	94.096%
	1/7	0.002	99.870%
	1/4	0.028	97.607%
	1/3	-0.004	100.316%
	1/2	-0.021	101.793%
1000	1/10	-0.021	101.838%
	1/7	-0.008	100.692%
	1/4	0.038	96.826%
	1/3	-0.022	101.874%
	1/2	0.034	97.160%
2000	1/10	0.013	98.956%
	1/7	-0.012	101.066%
	1/4	-0.001	100.046%
	1/3	0.013	98.841%
	1/2	-0.026	102.367%

Note. Biases that are greater than 0.05 or less than -0.05 are unacceptable.

Table A-5. Relative bias of ATE estimates and the percent bias reduction of ATE estimates with PSS.

Sample size	Simulated ratio of treated to total sample size	Relative bias of ATE estimates	Percent Bias Reduction of the ATE estimates
500	1/10	0.164	86.258%
	1/7	0.110	90.552%
	1/4	0.159	86.536%
	1/3	0.116	89.913%
	1/2	0.109	90.532%
1000	1/10	0.100	91.359%
	1/7	0.118	89.742%
	1/4	0.143	88.049%
	1/3	0.108	90.661%
	1/2	0.139	88.319%
2000	1/10	0.146	88.025%
	1/7	0.116	90.102%
	1/4	0.122	89.562%
	1/3	0.128	89.002%
	1/2	0.088	92.169%

Note. Biases that are greater than 0.05 or less than -0.05 are unacceptable.

APPENDIX B
RELATIVE BIAS OF STANDARD ERRORS OF THE ATE ESTIMATES

Table B-1. Relative bias of standard error estimates in the baseline.

Sample size	Simulated ratio of treated to total sample size	Empirical Standard Deviations	Relative bias of Weighted Least Squares standard error estimates	Relative bias of Taylor Series Linearization standard error estimates	Relative bias of Jackknife standard error estimates
500	1/10	28.903	-0.423	-0.042	-0.026
	1/7	24.435	-0.318	-0.040	-0.030
	1/4	19.107	-0.127	-0.002	0.003
	1/3	16.873	-0.011	0.045	0.048
	1/2	17.308	-0.034	-0.034	-0.032
1000	1/10	19.449	-0.393	-0.005	0.002
	1/7	17.153	-0.312	-0.029	-0.025
	1/4	13.310	-0.111	0.018	0.020
	1/3	12.244	-0.035	0.021	0.023
	1/2	11.835	-0.002	-0.002	-0.001
2000	1/10	13.853	-0.398	-0.011	-0.007
	1/7	11.893	-0.297	-0.005	-0.002
	1/4	9.553	-0.125	0.002	0.003
	1/3	8.842	-0.054	0.001	0.001
	1/2	8.535	-0.021	-0.021	-0.020

Note. Biases that are greater than 0.1 or less than -0.1 are unacceptable

Table B-2. Relative bias of standard error estimates with IPTW.

Sample size	Simulated ratio of treated to total sample size	Empirical Standard Deviations	Relative bias of Weighted Least Squares standard error estimates	Relative bias of Taylor Series Linearization standard error estimates	Relative bias of Jackknife standard error estimates
500	1/10	28.118	-0.407	0.058	0.086
	1/7	22.728	-0.267	0.089	0.105
	1/4	17.162	-0.026	0.154	0.161
	1/3	15.646	0.068	0.160	0.164
	1/2	15.641	0.071	0.095	0.097
1000	1/10	18.392	-0.358	0.113	0.125
	1/7	16.315	-0.276	0.068	0.075
	1/4	12.167	-0.026	0.150	0.154
	1/3	11.345	0.043	0.131	0.133
	1/2	10.786	0.098	0.119	0.120
2000	1/10	13.200	-0.367	0.092	0.098
	1/7	11.158	-0.250	0.109	0.112
	1/4	8.698	-0.037	0.134	0.136
	1/3	8.132	0.030	0.114	0.115
	1/2	7.612	0.100	0.119	0.120

Note. Biases that are greater than 0.1 or less than -0.1 are unacceptable

Table B-3. Relative bias of standard error estimates with TIPTW.

Sample size	Simulated ratio of treated to total sample size	Empirical Standard Deviations	Relative bias of Weighted Least Squares standard error estimates	Relative bias of Taylor Series Linearization standard error estimates	Relative bias of Jackknife standard error estimates
500	1/10	27.170	-0.387	0.060	0.083
	1/7	22.404	-0.258	0.082	0.095
	1/4	17.013	-0.019	0.150	0.157
	1/3	15.618	0.069	0.154	0.158
	1/2	15.671	0.068	0.088	0.091
1000	1/10	18.234	-0.354	0.092	0.102
	1/7	16.102	-0.268	0.062	0.068
	1/4	12.155	-0.027	0.141	0.144
	1/3	11.309	0.046	0.127	0.129
	1/2	10.783	0.097	0.116	0.117
2000	1/10	13.031	-0.361	0.078	0.083
	1/7	11.043	-0.244	0.100	0.103
	1/4	8.680	-0.037	0.126	0.128
	1/3	8.111	0.032	0.111	0.112
	1/2	7.609	0.100	0.117	0.117

Note. Relative biases that are greater than 0.1 or less than -0.1 are unacceptable

Table B-4. Relative bias of standard error estimates with PSS.

Sample size	Simulated ratio of treated to total sample size	Empirical Standard Deviations	Relative bias of Weighted Least Squares standard error estimates	Relative bias of Taylor Series Linearization standard error estimates	Relative bias of Jackknife standard error estimates
500	1/10	29.261	-0.463	0.046	0.080
	1/7	23.210	-0.324	0.082	0.100
	1/4	17.329	-0.093	0.142	0.149
	1/3	15.811	-0.009	0.146	0.150
	1/2	15.880	-0.011	0.075	0.077
1000	1/10	18.768	-0.410	0.096	0.109
	1/7	16.238	-0.320	0.076	0.084
	1/4	12.228	-0.097	0.141	0.145
	1/3	11.433	-0.036	0.118	0.120
	1/2	10.879	0.013	0.105	0.106
2000	1/10	13.169	-0.410	0.093	0.098
	1/7	11.160	-0.303	0.104	0.108
	1/4	8.698	-0.107	0.129	0.130
	1/3	8.112	-0.042	0.112	0.113
	1/2	7.639	0.015	0.110	0.111

Note. Relative biases that are greater than 0.1 or less than -0.1 are unacceptable

Table.B-5. Relative bias of standard error estimates with OFPSM.

Sample size	Simulated ratio of treated to total sample size	Empirical Standard Deviations	Relative bias of Weighted Least Squares standard error estimates	Relative bias of Taylor Series Linearization standard error estimates	Relative bias of Jackknife standard error estimates
500	1/10	28.527	-0.522	0.075	0.087
	1/7	24.278	-0.439	0.087	0.093
	1/4	18.574	-0.247	0.171	0.173
	1/3	17.407	-0.183	0.154	0.156
	1/2	17.179	-0.150	0.104	0.105
1000	1/10	19.932	-0.531	0.122	0.126
	1/7	17.622	-0.466	0.098	0.101
	1/4	13.499	-0.282	0.164	0.165
	1/3	12.308	-0.191	0.162	0.163
	1/2	11.662	-0.116	0.136	0.136
2000	1/10	14.687	-0.568	0.144	0.146
	1/7	12.419	-0.483	0.165	0.166
	1/4	9.687	-0.309	0.178	0.179
	1/3	8.769	-0.207	0.169	0.169
	1/2	8.006	-0.084	0.151	0.151

Note. Relative biases that are greater than 0.1 or less than -0.1 are unacceptable

APPENDIX C
EMPRICAL COVERAGE AND POWER TABLES

Table C-1. Empirical coverage rates of 95% confidence intervals across 1000 simulated data sets.

Sample size	Simulated ratio of treated to total sample size	Proportion coverage with IPTW	Proportion coverage with TIPTW	Proportion coverage with OFPSM	Proportion coverage with PSS
500	1/10	0.959	0.958	0.967	0.951
	1/7	0.953	0.954	0.962	0.957
	1/4	0.975	0.976	0.980	0.975
	1/3	0.979	0.975	0.983	0.970
	1/2	0.968	0.971	0.971	0.962
1000	1/10	0.969	0.956	0.975	0.961
	1/7	0.961	0.957	0.963	0.957
	1/4	0.977	0.975	0.969	0.976
	1/3	0.977	0.973	0.979	0.971
	1/2	0.971	0.970	0.974	0.961
2000	1/10	0.972	0.954	0.973	0.961
	1/7	0.962	0.962	0.977	0.956
	1/4	0.975	0.971	0.980	0.968
	1/3	0.977	0.972	0.981	0.966
	1/2	0.969	0.970	0.971	0.966

Note: Standard errors are estimated with Taylor Series Linearization.

Table C-2. Proportion of the estimated ATE that is significant at α=.05 level across 1000 simulated datasets.

Sample size	Simulated ratio of treated to total sample size	Proportion significant with IPTW	Proportion significant with TIPTW	Proportion significant with OFPSM	Proportion significant with PSS
500	1/10	0.110	0.132	0.085	0.118
	1/7	0.110	0.139	0.098	0.127
	1/4	0.153	0.173	0.123	0.187
	1/3	0.165	0.179	0.134	0.192
	1/2	0.190	0.203	0.161	0.248
1000	1/10	0.135	0.188	0.118	0.169
	1/7	0.200	0.240	0.150	0.229
	1/4	0.280	0.317	0.230	0.354
	1/3	0.314	0.348	0.250	0.393
	1/2	0.370	0.392	0.319	0.470
2000	1/10	0.280	0.365	0.194	0.339
	1/7	0.357	0.445	0.243	0.440
	1/4	0.530	0.585	0.404	0.636
	1/3	0.616	0.651	0.506	0.713
	1/2	0.635	0.664	0.579	0.759

Note: Standard errors are estimated with Taylor Series Linearization.

REFERENCES

Abadie, A. & Imbens, G. W. (2006). Large Sample properties of matching estimators for average treatment effects. *Econometrica*, *74*, 235-2667.

Abadie, A. & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76,* 1537-1557.

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

Austin, P. C. (2009a). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, *29*, 661-677.

Austin, P. C. (2009b). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics. 5(1),* Art. 13.

Austin, P. C. (2010a). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine, 29,* 2137-2148.

Austin, P. C. (2010b). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on propensity score. *Practice of Epidemiology, 172(9),*1092-1097.

Bembom, O., & van der Laan M. J. (2008). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *U.C. Berkeley Division of Biostatistics Working Paper Series.* Paper 230.

Cepeda M. S., Boston, R., Farrar, J. T., & Strom, B. L., (2003). Optimal matching with a variable number of controls vs. a fixed number of controls for a cohrot study: trade-offs. *Journal of Clinical Epidemiology, 56,* 230-237.

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

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

Cohen, J. (1973). Eta squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33, 107-112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Freedman, D. A. & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, *32(4),* 392-409.

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

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: statistical methods and applications.* Thousand Oaks: Sage

Hansen, B. B., & Fredrickson, M. (2009). optmach: Functions for optimal matching. Variable ratio, optimal, and full matching. Can also be implemented through MatchIt.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, 15(3), 234-349.

Heckman, J. J. (1978) Dummy endogenous variables in simultaneous equations system. *Econometrica*, *47*, 931-960.

Heckman, J. J., Ichimura, H,. & Todd, P. E. (1997) Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *65*, 261-294.

Hernan, M. A., Hernandez-Diaz, S. & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 82,* 387-394.

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

Hogan, J. W. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, *13*, 17-48.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research,* 26, 523-539.

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

Hsieah F. Y., Blooch D. A., & Larsen M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 17,* 1623-1634.

Lechner, M. (2008). A note on the common support problem in applied evaluation studies. *Annals of Economics and Statistics, Econometric Evaluation of Public Policies: Methods and Applications, 91/92,* 217-235.

Lohr, S. L. (1999). *Sampling: design and analysis.* Pacific Grove, CA: Duxbury Press.

Lumley, T. (2011). *"survey: analysis of complex survey samples".* R package version 3.62.1

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med., 23,* 2937-2960.

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*, 403-425.

McKelvey, R. D., & Zavoina,W. (1975).A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*, 103±120.

National Center for Education Statistics. (2010). School survey on crime and safety. Retrieved from http://nces.ed.gov/surveys/ssocs on June 1 2011.

Neugebauer, R., & van der Laan, M. (2005). Why prefer double robust estimates in causal inference? *Journal of Statistical Planning and Inference, 129,* 405-426.

Petersen, M. L., Wand, Y., van der Laan, M. J., & Bangsberg, D. R. (2006). Assessing the effectiveness of antiretroviral adherence interventions using marginal structural models to replicate findings of randomized controlled trials. *Acquir Immune Defic Syndr*, *43*, 96-103.

R Development Core Team. (2011). *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.* Retrived from http://www.Rproject.org.

Robins, J. M., Hernan, M. A. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550-560.

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

Rosenbaum, R. P. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistics Society*, *53*, 597-610.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

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

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

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34*, 441-456.

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

Rubin, D. B. (2007). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. *27*, 28-63.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Method, 13(4), 279-313.*

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

Stapleton, L. (2008). Chapter18: Analysis of data from complex surveys. In: E. D. de Leeuw, J. J. Hox & D. A. Dillman. *International handbook of survey methodology.* New York, Psychology Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods,* 15(3), 250-276.

Strayhorn, T. L. (2009). Accessing and analyzing national databases. In T. J. Kowalski & T. J. Lasley II (Eds.), Handbook of data-based decision making in education (pp. 105-122). New York, NY: Routledge.

Strumer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution-a simulation study. *Practice of Epidemiology, 172(7),* 842-854.

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

Thoemmes, F. J., Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46,* 90-118.

Thoemmes, F., & West, S. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46,* 514-543.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wang, Y., Petersen, M. L., Bangsberg, D., & van der Laan M. J. (2006). Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. *U.C. Berkeley Division of Biostatistics Working Paper Series,* Working Paper 211.

Winship, C. & Morgan, S. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659-706.

Wolter, K. M. (2007). *Introduction to Variance Estimation.* New York: Springer.

BIOGRAPHICAL SKETCH

Sungur Gurel was born in Osmaniye, Turkey. He received his B.A. in Mathematics Teaching from Istanbul University at 2003. He served for the Turkish Government as a Mathematics teacher for 20 months. He later qualified for a scholarship to study abroad in Fall 2010, enrolled for graduate studies in the School of Human Development and Organizational Studies at Collage of Education at the University of Florida. He has been awarded an Honorable Mention award by Educational Statisticians Special Interest Group of American Educational Research Association in April 2012. He will receive his M.A.E in Research and Evaluation Methodology program in August 2012.