


January 2013

Effectiveness of Propensity Score Methods in a Multilevel Framework: A Monte Carlo Study

Aarti P. Bellara

University of South Florida, aartibellara@gmail.com

Follow this and additional works at: <http://scholarcommons.usf.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Statistics and Probability Commons](#)

Scholar Commons Citation

Bellara, Aarti P., "Effectiveness of Propensity Score Methods in a Multilevel Framework: A Monte Carlo Study" (2013). *Graduate Theses and Dissertations*.

<http://scholarcommons.usf.edu/etd/4635>

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

Effectiveness Of Propensity Score Methods In A Multilevel Framework:

A Monte Carlo Study

by

Aarti P. Bellara

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Educational Measurement and Research
College of Education
University of South Florida

Major Professor: Jeffrey D. Kromrey, Ph.D.
John M. Ferron, Ph.D.
Eun Sook Kim, Ph.D.
Zorka Karanxha, Ed.D.

Date of Approval:
May 15, 2013

Keywords: educational evaluations, hierarchical linear modeling, observational studies, causality,
simulation research

Copyright © 2013, Aarti P. Bellara

DEDICATION

I would like to dedicate this dissertation to several very important people. First, this dissertation is dedicated in loving memory of my grandparents. I am forever indebted to you for blessing me with such amazing parents. May you continue to rest in peace and as you look down on me, feel proud to call me your granddaughter.

Second, this dissertation is dedicated to my parents, Partab and Hiru Bellara. There are simply not enough words to truly capture my emotions and gratitude to them; thank you seems so insufficient for all they have done. Nothing could have been accomplished without their unconditional love, support, and prayers. The sacrifices and struggles they have made for our family have not gone unnoticed. Thank you for being the biggest fan of my life. I hope I have made you proud.

Finally, this dissertation is dedicated to my brother, Amar Bellara. Thank you for teaching me how to take the challenges of life and turn them into fun, and for always being my best friend.

What can you do to promote world peace? Go home and love your family.

Mother Teresa

ACKNOWLEDGMENTS

First and foremost, I thank God for all the blessings in my life. Thank you for providing me with the perseverance, courage, and wisdom to embark and complete this journey.

The creation and completion of this dissertation would have occurred without the support of my doctoral committee. My major professor and mentor, Dr. Jeffrey D. Kromrey, provided me with the guidance and support I needed to envision, conduct, and complete this study. It was his supportive and encouraging nature that helped me make the decision to pursue my doctoral studies at the University of South Florida, and it was he who mentored me through the very end. Thank you for taking me under your wing, sculpting me into the scholar I have become, and helping me realize my own potential.

Heartfelt thanks to the other members of my doctoral committee. Dr. John M. Ferron stuck by my side throughout my entire doctoral academic career, and constantly challenged me to find my own path. He is an outstanding professor and mentor and I strive to emulate his philosophy to teaching and research. Dr. Eun Sook Kim, joined me as I began this dissertation and not only supported me, but also challenged me to continue to learn and grow through this process. Your expertise in the field was an invaluable asset to me. Dr. Zorka Karanxha, provided me with my first research experience and taught me countless lessons throughout this process, including how to be an accomplished, respected, and compassionate researcher.

In addition to my doctoral committee, the whole Department of Educational Measurement and Research has impacted my journey through doctoral studies from our office manager, Jody Duke, to the rest of the faculty, both current and former (Drs. Constance Hines, Christopher DeLuca, Liliana Rodriguez-Campos, Robert Dedrick, Yi-Hsin Chen, Jeanine Romano, and Jennifer Wolgemuth). Several students and graduates have played a major role

throughout my doctoral journey; we started off as classmates and continue to be colleagues and friends: Susan Hibbard, Merlande Petit-Bois, Elly Baek, Connie Walker, Corina Owens, Heather Scott, Jennie Farmer, and Bethany Bell. Special thanks to Patricia Rodriguez de Gil, Diep Nguyen, Thanh Pham, for all your help and support during my data collection phase.

Without the loving support of family, nothing would be possible. Thanks to my parents for teaching me about the value of education and sacrificing so much so that I could accomplish my goals, to my aunt, Dr. Jyoti Chandiramani for being my mentor, friend, and academic colleague, and to the rest of my extended family and friend for providing encouragement throughout my journey. Lastly, from the bottom of my heart, thanks to those within the Indian community of the greater Tampa Bay area who took care of me and made me a part of your own families throughout the past five years.

A small body of determined spirits fired by an unquenchable faith in their mission can alter the course of history.

Mahatma Gandhi

TABLE OF CONTENTS

List of Tables	iv
List of Figures	vi
Abstract	viii
Chapter One: Introduction	1
Causal Inference	1
Propensity Scores	3
Problem Statement	5
Study Purpose	7
Research Questions	8
Overview of the Study	8
Delimitations	11
Significance of the Study	11
Limitations	13
Definition of Terms	13
Chapter Two: Literature Review	16
Theoretical Framework	16
Rubin's Causal Model	17
Strongly ignorable treatment assignment assumption	18
Stable unit treatment value assumption	19
Campbell's Validity Framework	20
Causal Inference in Non-Randomized Studies	22
The Logic of Propensity Scores	23
Covariate Selection	24
Estimation Methods	24
Conditioning Methods	26
Matching	27
Stratification	28
Covariance adjustment	29
Weighting	30
Evaluating the Accuracy of the Propensity Score Model	31
Practical Concerns with Propensity Score Analysis	33
Covariate Selection	33
Estimation Methods	40
Conditioning Methods	45
Evaluating the Accuracy of the Propensity Score Models	48
The Overall Effectiveness of Propensity Score Methods	50
Multilevel Modeling	54
Propensity Score Analysis with Multilevel Modeling	56

Multilevel research designs	56
Estimation Models	57
Conditioning Methods	59
Research on Propensity Scores in Multilevel Contexts	60
Chapter Summary	67
Chapter Three: Method	68
Study Purpose	68
Research Questions	68
Design	69
Samples	70
Sample Characteristics	73
Sample size	73
Relationship between covariates and treatment assignment	74
Relationship between covariates and outcome	76
Population treatment effect	77
Analytical Procedures	77
Step one: Propensity score estimation	79
Step two: Evaluating the region of common support	85
Step three: Propensity score conditioning	86
Step four: Assessment of balance	89
Step five: Estimate treatment effects	92
Step six: Final comparative analysis	93
Chapter Summary	94
Chapter Four: Results	95
Overview of the Study	95
Description of Samples	96
Propensity Score Estimation Models	96
Common Support	101
Propensity Score Conditioning	104
Data Analysis	108
Balance	110
Treatment Effects	122
Answers to Research Questions	135
Research Question 1: To what extent do balance estimates vary across PS methods (PS estimation models and PS conditioning strategies)?	135
Research Question 2: To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the balance achieved by the PS methods (PS estimation models and PS conditioning strategies)?	136
Research Question 3: To what extent do treatment effect estimates vary across PS methods (PS estimation models and PS conditioning strategies)?	137
Research Question 4: To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the treatment effects achieved by the PS methods (PS estimation models and PS conditioning strategies)?	138

Research Question 5: What is the direction and strength of the relationships between balance and both the accuracy and precision of the treatment effect estimates?.....	139
Chapter Summary	141
Chapter Five: Discussion	142
Summary of the Study	142
Purpose	142
Research Questions.....	142
Method.....	143
Discussion of the Study Results.....	144
Balance	144
Treatment Effects.....	146
The Relationship Between Balance and the Accuracy and Precision of Treatment Effects.....	149
Limitations of the Study	150
Implications	153
Implications for Researchers Conducting PS Analysis with MLM.....	153
Implications for Methodologists	154
References.....	156
Appendices	170
Appendix A: Equation for Data Generation	170
Appendix B: Population R^2 Values Simulated	173
Appendix C: Equations for Propensity Score Estimation.....	174

LIST OF TABLES

Table 1:	Design Features	10
Table 2:	Conditioning Methods	11
Table 3:	Covariate Relationship to Assignment and Outcome	35
Table 4:	Data structure for Lee and associates (2010) simulation study	42
Table 5:	Average Distributions of the Propensity Scores for each Estimation Model for Treatment and Control Groups.....	96
Table 6:	Descriptive Statistics for Mean Correlations between Propensity Score Estimates	97
Table 7:	Descriptive Statistics for Mean Non-Positive Definite Matrix Rates for each Multilevel PS Estimation Model by Level 1 Sample Size.....	101
Table 8:	Mean Propensity Score Range Before and After Trimming.....	101
Table 9:	Mean Percent of Data Trimmed for Covariate Relationship to Treatment by PS Model	104
Table 10:	Mean Proportion of Potential Matches for Covariate Relationship to Treatment by PS Model	107
Table 11:	Descriptive Statistics by Design Factors Associated with Mean Balance Score	112
Table 12:	Mean Number of Unbalanced Covariates by Level-1 Sample Size and Conditioning Method.....	116
Table 13:	Descriptive Statistics for the Mean Number of Unbalanced Covariates by Level-2 Sample Size	118
Table 14:	Descriptive Statistics for the Proportions of Samples Balanced by Level 1 Sample Size, Level-2 Sample Size and Conditioning Method.....	122
Table 15:	Descriptive Statistics by Design Factors Associated with Bias.....	124
Table 16:	Root Mean Squared Error by Level 1 Sample Size Across the Clusters	127

Table 17:	Confidence Interval Coverage by Level 1 Sample Size Across the Clusters	130
Table 18:	Confidence Interval Coverage Estimates by Conditioning Method, Estimation Model and Level-2 Sample Size.....	132
Table 19:	Confidence Interval Width Averages and Distributions By Sample Size Interaction Effect Across Propensity Score Models	135
Table 20:	Correlations between balance score and absolute value of bias estimates for PS estimation models across PS conditioning methods.....	139
Table 21:	Correlations between balance score and RMSE estimates for PS estimation models across PS conditioning methods.....	140
Table 22:	Correlations between balance score and confidence interval coverage estimates for PS estimation models across PS conditioning methods	141
Table 23:	Correlations between balance score and confidence interval width for PS estimation models across PS conditioning methods	141
Table A1:	R^2 values for the different confounder magnitudes	173

LIST OF FIGURES

Figure 1:	Propensity score methodology framework	5
Figure 2:	Analytical procedures with corresponding outcome measures.....	78
Figure 3:	Mean non-convergence distributions by propensity score estimation model	98
Figure 4:	Mean non-convergence distributions by level 1 sample size across the number of clusters for each PS model	99
Figure 5:	Mean non-positive definite matrix rates by propensity score estimation model	100
Figure 6:	Mean percent of data trimmed by propensity score estimation model	102
Figure 7:	Mean Percent of Data Trimmed by level-1 sample size across level 2- sample size for each propensity score model.....	103
Figure 8:	Distributions of mean proportion of potential matches across PS models	106
Figure 9:	Proportion of samples dropping strata by propensity score estimation model	107
Figure 10:	Proportion of Samples Dropping Strata by the Sample Size Interaction across the Four Propensity Score Models.....	109
Figure 11:	Absolute average standardized mean difference in covariates by conditioning method across the four PS models.....	111
Figure 12:	Distributions of the absolute average standardized mean difference in covariates by conditioning method across the four PS models for each level-1 sample size level.....	113
Figure 13:	Absolute average standardized mean difference in covariates by conditioning method across the four PS models for each level-2 sample size level	114
Figure 14:	Mean number of unbalanced covariates by conditioning method across the four PS models.....	115
Figure 15:	Distributions in mean number of unbalanced covariates by level-1 sample size across conditioning methods	116

Figure 16:	Distributions in mean number of unbalanced covariates by level-1 sample size across conditioning methods by propensity score estimation model.....	117
Figure 17:	Distributions of mean number of unbalanced covariates by conditioning method across the four PS models for each level-2 sample size level.....	119
Figure 18:	Proportion of samples balanced by conditioning method across PS models.....	120
Figure 19:	Proportion of samples balanced by conditioning method and level-1 sample size interaction across the level-2 sample size	121
Figure 20:	Distributions of estimated bias in point estimates by conditioning methods across PS models.....	123
Figure 21:	Distributions of the average bias in point estimates by conditioning method across the four PS models for each level-1 sample size level.....	125
Figure 22:	Distributions of the RMSE for each conditioning method across the PS estimation models	126
Figure 23:	Root mean squared error for the level 1 sample size across the number of clusters	127
Figure 24:	Distributions of 95% confidence interval coverage rates for each conditioning method across the four PS estimation models.	128
Figure 25:	Distributions of 95% confidence interval coverage rates by level-1 sample size across the number of clusters	129
Figure 26:	Distributions of 95% confidence interval coverage rates by PS estimation models and conditioning methods across the number of clusters.....	131
Figure 27:	Distributions of confidence interval width by conditioning methods across PS estimation models.....	132
Figure 28:	Distributions of the confidence interval width by propensity score estimation model and level-1 sample size interaction across the level-2 sample size.....	134

ABSTRACT

Propensity score analysis has been used to minimize the selection bias in observational studies to identify causal relationships. A propensity score is an estimate of an individual's probability of being placed in a treatment group given a set of covariates. Propensity score analysis aims to use the estimate to create balanced groups, akin to a randomized experiment. This study used Monte Carlo methods to examine the appropriateness of using propensity score methods to achieve balance between groups on observed covariates and reproduce treatment effect estimates in multilevel studies. Specifically, this study examined the extent to which four different propensity score estimation models and three different propensity score conditioning methods produced balanced samples and reproduced the treatment effects with clustered data. One single-level logistic model and three multilevel models were investigated. Conditioning methods included: (a) covariance adjustment, (b) matching, and (c) stratification. Design factors investigated included: (a) level-1 sample size, (b) level-2 sample size, (c) level-1 covariate relationship to treatment, (d) level-2 covariate relationship to treatment, (e) level-1 covariate relationship to outcome, (f) level-2 covariate relationship to outcome, and (g) population effect size. The results of this study suggest the degree to which propensity score analyses are able to create balanced groups and reproduce treatment effect estimates with clustered data is largely dependent upon the propensity score estimation model and conditioning method selected. Overall, the single-level logistic and random intercepts models fared slightly better than the more complex multilevel models while covariance adjustment and matching methods tended to be more stable in terms of balancing groups than stratification. Additionally,

the results indicate propensity score analysis should not be conducted with small samples. Finally, this study did not identify an estimation model or conditioning method that was consistently able to create adequately balanced groups and reproduce treatment effect estimates.

CHAPTER ONE: INTRODUCTION

Causal Inference

Identifying causal relationships in social settings has been and continues to be a challenge. Educational researchers often seek to identify causal relationships between programs, interventions, and/or treatments (herein "treatments") on various student outcomes, such as academic achievement (Austin, 2011; Chatterji, 2008; Slavin, 2002, 2008). Experiments investigate treatment effects of manipulable causes using statistical models to draw causal inferences. Manipulable causes are ones that can be deliberately altered or manipulated by the researcher, for example, participation in a program, method of teaching, or dosage amount. In contrast, non-manipulable agents such as gender cannot be deliberately altered and therefore, are not causes in experiments (Shadish, 2010; Shadish, Cook, & Campbell, 2002). Consequently, identifying causal relationships among non-manipulable variables becomes challenging (Shadish et al., 2002).

Causal relationships exist between two variables when the following hold true: (a) the cause precedes the effect, (b) the cause is related to the effect, and (c) no plausible alternative explanations for the effect exist other than the cause (Shadish et al., 2002). Treatment effects are estimated by a counterfactual model, or simply, the difference between what did happen after an individual received a treatment versus what would have happened if the same individual did not receive the treatment (Campbell & Stanley, 1963; Holland, 1986; Rubin, 2010; Shadish et al., 2002). Theoretically, causal effects can be precisely estimated if a unit was assigned to the treatment condition and the control condition at the same time in the same context. This would

allow outcome values for each unit under both of the conditions to be observed (Rubin, 1974, 1978).

In most field based settings, units can be assigned to one condition; therefore, only one outcome is observed—the outcome of the condition to which the individual was assigned. The unobserved or missing outcome is considered the counterfactual. For example, to investigate the effectiveness of a new reading program, an individual cannot be assigned to the new reading program and the old reading program simultaneously. The impossibility of observing both treatment and control outcomes for each individual is often referred to as the "Fundamental Problem of Causal Inference" (Holland, 1986, p. 947; Rubin, 1978).

Randomized controlled trials (RCTs), or randomized experiments, are considered the "gold standard" for estimating treatment effects (Austin, 2011; Cook, 2006; Donaldson & Christie, 2004; Education Sciences Reform Act [ESRA], 2002; Scriven, 2008; United States Department of Education [USDOE], 2003, 2005). In a randomized experiment, individuals are randomly assigned to treatment conditions. Random assignment allows groups to be probabilistically similar, supporting a counterfactual inference; therefore, any measured differences in the outcome may be attributed to treatment effect (Campbell & Stanley, 1963; Games, 1990; Holland, 1986; Shadish, et al., 2002).

The process of randomization guarantees the two groups, on average, will be balanced at the beginning of the experiment, except for treatment assignment, and thus able to yield estimates of the average treatment effect (ATE). Estimates are considered to be unbiased because the randomization process ensures no plausible alternative explanation exists. Accordingly, well executed RCTs are able to produce unbiased estimates of treatment effects and are, therefore, the preferred method when investigating causal relationships. However, random assignment is often unethical or impractical in social and behavioral research. For example, to investigate the effectiveness of private school education on student achievement, the researcher is generally

unable to randomly assign students to private (treatment) or public (control) schools.

Consequently, non-randomized studies are often used to estimate treatment effects (Austin, 2011).

In contrast, experiments which do not employ random assignment techniques, yet aim to explore causation, provide “less compelling support for counterfactual inferences” (Shadish, et al, 2002, p. 14) because groups are not probabilistically similar. In addition, causal relationships from non-manipulable variables may also be identified. Rubin (1979, 2001, 2007, & 2008) uses the term "observational studies" to refer to all studies aiming to explore causal relationship that do not incorporate the randomization process. Both non-randomized-experiments and observational studies (herein non-randomized studies) lack the desired properties for causal inference and subsequently the validity of inferences are subject to various threats. These threats introduce sources of alternative explanations of the treatment effects which are thus considered to be potentially biased and less precise (Campbell & Stanley, 1963; Shadish, 2000; Shadish, et. al., 2002).

There are two basic approaches to addressing the estimation of causal effects in non-randomized studies: alternative design features (e.g. regression discontinuity, interrupted time-series) and applied statistical methods (e.g. ordinary regression, covariance adjustment analysis, structural equation modeling, selection models, and matching methods) (Gall, Gall, & Borg, 2007; Shadish, 2000; Shadish, et al., 2002; Stuart, 2010). Throughout the 20th century, methodologists have worked to develop, refine, and evaluate these approaches. One of the more recent developments, introduced by Rosenbaum and Rubin (1983) is propensity score adjustment, a statistical method which aims to achieve balance between groups on a set of observed covariates with a single number (Stuart, 2010).

Propensity Scores

A propensity score (PS) is the “conditional probability of assignment to a particular group, given a vector of covariates” (Rosenbaum & Rubin, 1983 p. 42). The purpose of the PS is to improve the quality of estimates from non-randomized experiments by attempting to mimic the balance between groups that occurs through the randomization process (Rosenbaum & Rubin, 1984; Shadish & Steiner, 2010; Stuart, 2010). The PS predicts an individual's probability for being assigned to the treatment group, thus ranges from 0 to 1. The closer the individual's PS is to 1, the stronger the prediction for being in the treatment group; conversely, the closer the score is to 0, the stronger the prediction for being in the comparison group. When units from the treatment and control group have the same propensity score, it is assumed that the probability of being assigned to the treatment group is the same for each of these individual units, conditional upon the observed covariates. When there is no overlap in PSs between the groups, it is believed that the unobserved covariate(s) are accounting for the difference in groups (Stuart, 2010).

In randomized experiments, when treatment and control group samples have the same number of participants, the probability of being placed in treatment or control is equal (each participant's PS = .50), and the two groups are considered to be comparable with differences being attributable to chance (Shadish & Steiner, 2010; Steiner & Cook, 2013). When randomization is not possible, the probability of being placed in treatment or control is unknown but can be estimated. Equation 1 represents Rosenbaum and Rubin's (1983) formulation to represent an individual, i 's probability of receiving the treatment, $P(Z_i=1)$, given a set of observed covariates, X . The resulting probability, $e(X)$, is the estimated PS:

$$e_i(X_i) = P(Z_i = 1 | X_i) \tag{1}$$

Estimating treatment effects using PSs is a multistep decision-based process that commonly includes the following procedures: (a) selecting the appropriate covariates related to both assignment and treatment to include in the model, (b) estimating the PSs, (c) conditioning on

the PS, (d) assessing the accuracy of the PS estimation model, (e) adjusting the model if necessary, and finally, (f) estimating treatment effects (Rosenbaum & Rubin, 1983; Shadish & Steiner, 2010; Stuart, 2010) (see Figure 1).

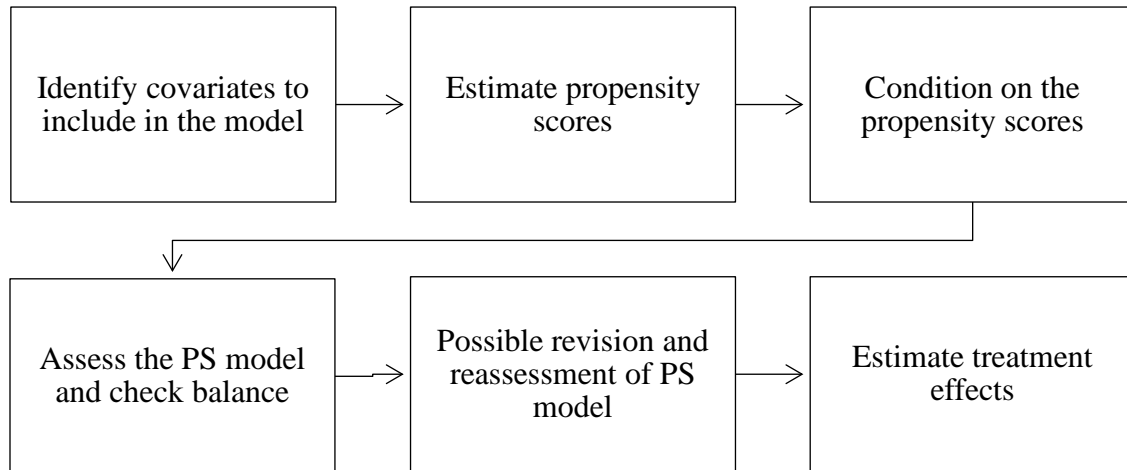


Figure 1. Propensity score methodology framework.
Note: Figure is derived from Thoemmes, 2009, p. 150

Since Rosenbaum and Rubin’s seminal work in 1983, the study and application of PS has grown in popularity. Researchers have investigated the performance of PS techniques using Monte Carlo simulation studies (e.g. Austin, 2009; Gu & Rosenbaum, 1993) as well as existing databases (e.g. Michaelopoulos, Bloom, & Hill, 2004; Stuart & Green, 2008). Additionally, PSs have been applied to non-randomized studies across various fields including medicine (e.g. Murphy, Law, Whooley, Alexandrou, Chu, & Wong, 2003), business (e.g., Dehejia & Wahba, 1999), and social sciences (e.g. Hong & Yu, 2008).

Problem Statement

While RCTs may yield unbiased estimates of treatment effects, they are often difficult to implement in educational settings. Therefore, causal relationships in educational research are likely to be drawn from non-randomized studies which lack the desired properties to support counterfactual inferences and are highly vulnerable to threats to validity (Shadish et al., 2002).

Because PSs aim to mimic certain characteristics of random assignment, their application is rapidly increasing. Historically, PS analysis has been used extensively in the medical field; however, recently it has begun to receive attention in other fields (Hahs-Vaugh & Onwuegbuzie, 2006; Pruzek, 2011; Thoemmes & Kim, 2011). Weitzen and colleagues (2004) investigated the PS estimation models considered by medical researchers, limiting their analysis to studies published during 2001 only. A total of 47 studies were ultimately included in their analysis. Similarly, Austin (2008a) also reviewed 47 studies published within the medical literature during 1996-2003. Austin (2008a) limited his analysis to applications employing only one conditioning method, matching. Reviews of PS applications in the social sciences have reported comparable numbers; however, these reviews did not impose strict criteria for inclusion and included a wider range of years across multiple disciplines. For example, Hahs-Vaugh & Onwuegbuzie (2006) searched the Education Resources Information Center (ERIC) database for educational studies applying PS techniques and identified a total of 25 applications of PS in education. Similarly, Thoemmes and Kim (2011) searched three large databases (ERIC, PsycINFO, and Web of Science), for PS applications in social science fields. Their search generated 111 studies published from 1991-2008, of which, 86 met their inclusion criteria, and 34 were from education.

The gradual adoption of PSs in the social sciences may be associated with the fact that there still remains a plethora of research situations for which there is a paucity of empirical evidence justifying the appropriateness of PSs (Shadish & Steiner, 2010). For example, investigations of the behavior and application of PSs with nested data has received relatively little attention. Much of the research on PSs assumes observations to be independent (Arpino & Mealli, 2011; Hahs-Vaughn & Onwuegbuzie, 2006; Lingle, 2009); however, many systems, especially those in education, include a hierarchical structure where individuals are nested in multiple levels (e.g., students nested within classrooms; classrooms nested within schools; schools nested within districts) a violation of the independence assumption.

Hierarchical linear modeling (HLM), also referred to as multilevel modeling (MLM), is often employed when analyzing data with a nested or hierarchical structure (Raudenbush & Bryk, 2002). Estimating treatment effects within a nested context is complex because outcomes often depend upon contextual factors of the various levels. Additional challenges for drawing causal inferences within a MLM framework include implications associated with different hierarchical designs (Hong & Raudenbush, 2003). Currently, there is little guidance provided on how to incorporate statistical methods, such as PSs to draw causal inferences with nested structures (Thoemmes & West, 2011).

Recently, PS methods have been applied to studies conducted in multilevel settings (e.g., Hong & Raudenbush, 2005; Hong & Yu, 2008; Kim & Seltzer, 2007). Additionally, few studies have also recently begun to evaluate the performance of different PS methods with hierarchical data (Arpino & Mealli, 2011; Lingle, 2009; Rodriguez de Gil et al., 2012; Thoemmes, 2009; Thoemmes & West, 2011). While these studies have contributed to the body of knowledge, the literature on PSs with MLM remains sparse. Given the steady increase of PSs, coupled with the desire to investigate causal relationships in educational settings, further examination of the performance of PSs with MLM is warranted and timely.

Study Purpose

The purpose of this study was to further examine the appropriateness of using PS methods to achieve balance between groups on observed covariates and to yield unbiased treatment effect estimates in multilevel studies. Specifically, this study examined the extent to which different PS approaches (PS estimation models and PS conditioning techniques) and sample characteristics (sample size, covariate relationship to treatment and outcome, and population effect size) achieved balance and reproduced the population treatment effect.

PSs were estimated using four different logit models (a) single level model, (b) fixed slopes with random intercepts ignoring cluster-level predictors (random intercepts), (c) random

slopes and intercepts ignoring the cluster-level predictors (random coefficients) and, (d) random slopes and intercepts with cluster-level predictors added (cross-level). For each of the four PS estimation models, three different PS conditioning strategies were also investigated and included: (a) matching, (b) stratification, and (c) covariance adjustment. PS methods (estimation models and conditioning strategies) fully crossed with the sample characteristics were examined to evaluate the quality of balance achieved as well as the accuracy and precision of treatment effect estimates produced.

Research Questions

1. To what extent do balance estimates vary across PS methods (PS estimation models and PS conditioning strategies)?
2. To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the balance achieved by the PS methods (PS estimation models and PS conditioning strategies)?
3. To what extent do treatment effect estimates vary across PS methods (PS estimation models and PS conditioning strategies)?
4. To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the treatment effects estimated by the PS methods (PS estimation models and PS conditioning strategies)?
5. What is the direction and strength of the relationship between balance and both the accuracy and precision of treatment effect estimates?

Overview of the Study

This study incorporated Monte Carlo simulation methods to examine the performance of PS methods with MLM. Simulation methods allow for the control and manipulation of specific design and data factors to investigate the behavior of statistical methods (Guo & Fraser, 2010). This current study included nine design factors (see Table 1) related to either PS technique or

sample characteristics. These factors are (a) PS estimation models (single level, random intercepts, random coefficients, and cross-level); (b) PS conditioning strategies (matching, stratification, and covariance adjustment); (c) number of clusters (small [n=30], moderate [n=50], and large [n=100]); (d) within-cluster sample size (small [n =01-09], moderate [n =10-19], and large [n =20-29]); (e) relationship between level-1 covariates and treatment assignment (small [$\beta_{xz}=.10$], and moderate [$\beta_{xz}=.20$]); (f) relationship between level-1 covariates and outcome (small [$\beta_{xy}=.10$], and moderate [$\beta_{xy}=.20$]); (g) relationship between level-2 covariates and treatment assignment (small [$\gamma_{wz}=.20$], and moderate [$\gamma_{wz}=.40$]); (h) relationship between level-2 covariates and outcome (small [$\gamma_{wy}=.20$], and moderate [$\gamma_{wy}=.40$]); and (i) population effect size (δ = small [0.2] and moderate [0.5]). All levels of all the factors were fully crossed with one another for a total of 288 data conditions (see Table 1). For each of the 288 data conditions, twelve different combinations of propensity score methods were conducted yielding a total of 3,456 conditions. For 254 out of 288 data conditions, 1000 datasets were simulated using SAS IML (SAS Institute Inc., 2008). The remaining 36 conditions, where the number of clusters was 100 and the number of level 1 units within these 100 clusters was large (20-29), 500 datasets were simulated using SAS IML (SAS Institute Inc., 2008).

Two specific aspects of PS methodology within a MLM framework were of interest in this study: the quality of the balance achieved and the accuracy and precision of the treatment effect estimates. The standardized differences for the estimated PSs and the observed covariates after conditioning were used to estimate balance. Outcome measures associated with effective treatment effect estimates included bias, standard error, 95% confidence interval coverage and width.

This study incorporates multiple analytical steps. First, PSs were estimated for each of the four models. Next, the PS estimates and samples were evaluated and trimmed to include only the common support areas. Then, each of the four PS model's trimmed samples were conditioned

Table 1
Design Features

Sample Characteristics							PS Methods			
							PS estimation models			
							Single level	Random intercepts	Random-coefficients	Cross-level
							PS conditioning strategies*			
Sample size		Covariate relationship to treatment		Covariate relationship to outcome		Population effect size	M	M	M	M
Number of clusters	Sample size within cluster	Level 2	Level 1	Level 2	Level 1		S	S	S	S
							C	C	C	C
30	01-09	$\gamma_{os}=.20$	$\beta=.10$	$\gamma_{os}=.20$	$\beta=.10$	$\delta=.2$				
	10-19	$\gamma_{os}=.40$	$\beta=.20$	$\gamma_{os}=.40$	$\beta=.20$	$\delta=.5$				
	20-29									
50	01-09	$\gamma_{os}=.20$	$\beta=.10$	$\gamma_{os}=.20$	$\beta=.10$	$\delta=.2$				
	10-19	$\gamma_{os}=.40$	$\beta=.20$	$\gamma_{os}=.40$	$\beta=.20$	$\delta=.5$				
	20-29									
100	01-09	$\gamma_{os}=.20$	$\beta=.10$	$\gamma_{os}=.20$	$\beta=.10$	$\delta=.2$				
	10-19	$\gamma_{os}=.40$	$\beta=.20$	$\gamma_{os}=.40$	$\beta=.20$	$\delta=.5$				
	20-29									

M= Matching
S= Stratification
C= Covariance Adjustment

three different times. Lastly, treatment effects were estimated using MLM, and balance on the conditioned samples was assessed. General Linear Models (GLM) procedures were conducted to address the research questions and draw inferences about the variability in balance and treatment effect estimates across PS methods and sample characteristics.

Delimitations

In addition to the manipulated factors, several design factors were held constant throughout this study. Data simulated included 27 continuous and 3 dichotomous level 1 covariates (X), 9 continuous and 1 dichotomous level 2 covariates (W), 1 binary assignment variable (Z) and 1 continuous outcome variable (Y). Data were generated so that the correlation between covariates within each level is approximately 0.2. Specific details regarding the three conditioning methods are presented in Table 2.

Table 2

Conditioning Methods

1. Matching - 1:1 Nearest neighbor caliper matching without replacement
 2. Stratification- Five strata distributed evenly
 3. Covariance Adjustment-including the estimated PS as varying level-1 covariate
-

Significance of the Study

Since the early 80's researchers have studied PS methods as well as multilevel models with respect to their ability to estimate effectiveness, individually. Many social and behavioral studies employ non-randomized studies in hierarchical settings to estimate treatment effects. Given that non-randomized studies lack the desirable properties of RCTs, it is imperative for methodologists to examine and improve methods used to identify causal relationships. The PS is a fairly recent statistical approach, and examinations of this method within MLM have received little attention. By examining the performance of PS in MLM, this study aimed to contribute to an important gap in the ongoing dialogue on causal inference in social and behavioral field settings.

Although recently researchers have begun to investigate the behavior of PSs with MLM through simulation studies, these studies have not considered conditions often found in applied educational settings. For example, previous studies have included relatively few level 1 predictors (e.g. less than 10) to balance groups (e.g. Arpino & Mealli, 2011; Lingle, 2009; Thoemmes & West 2011). In their review of PS applications in social science fields, Thoemmes and Kim (2011) reported an average of 31.3 covariates used in 79 studies. Inferences drawn based on investigations using ten covariates should be considered tentative when generalizing results to applications using more than ten.

The simulated samples in this study aimed to represent current applications of PS analyses in social science research and common attributes of educational data, specifically, focusing on research situations where units nested in clusters are assigned to treatment group with conditioning occurring across clusters. Several aspects distinguish the current study from previous ones, particularly in sample complexity and PS methods investigated. Specifically the following characteristics induce a degree of complexity to the samples and have not previously been investigated within this context: (a) larger number of level-1 and level-2 predictors, (b) correlations among the predictors, and (c) dichotomous covariates. In addition, this study investigated a broad range of PS methods. For example, three different PS conditioning techniques were examined in order to introduce and extend additional PS methods to a multilevel framework. Currently, little empirical evidence exists on which PS conditioning method work best under certain situations. Only a few studies compare the conditioning techniques using single level, non-nested data. No previous investigation of PS methods comparing three different conditioning techniques using MLM could be located. Lastly, the majority of methodological studies of PSs focus on the degree to which the PSs are able to remove bias in treatment effects; balance estimates are often a secondary outcome if and when included. Consequently, balance estimates are not consistently reported in applied studies incorporating PS methods (Thoemmes,

& Kim, 2011). Balance indicates the PSs' ability to create comparable groups that would mimic a random sampling design, an important, albeit overlooked purpose of PS models.

The findings from this study build upon the current literature and offer multiple avenues for future research. While this study's significant contributions are primarily methodological, current trends in social and behavioral research suggest the timely nature and opportunity for this study to provide applied researchers across disciplines additional information on the nature of causal inference with non-randomized nested data.

Limitations

Although this study contributes to the methodological literature on causal inference of non-randomized studies, specifically the use of propensity scores in multilevel contexts it is not without limitations. There are numerous combinations and permutations of potential design factors that may be considered. Data in this study was based on the specific aforementioned conditions and delimitations. While simulation studies are intended to provide evidence and rationale for empirical application, findings from this study can only be generalized to studies with similar conditions.

Definition of Terms

Assignment- binary variable that determines whether an individual or cluster receives treatment or control.

Balance- met when the distribution of observed covariates is equal between treatment and control groups.

Bias- the difference between a known parameter estimate and the estimated parameter estimate

Common Support- the region of overlap in estimated PSs across treatment groups

Conditioning Strategies- methods employed to apply the propensity scores to balance the groups on the observed covariates.

Confidence Interval Coverage- the proportion of 95% confidence intervals that includes the estimated parameter.

Confidence Interval Width- the difference between the upper and lower limits of the 95% confidence intervals for the estimated parameter. This statistic will be aggregated across replications within each condition and represent the Average confidence interval width.

Confounding variable- a variable that is related to both treatment assignment and outcome.

Control- the group not receiving or exposed to a specific condition under investigation of its effectiveness. This group is often known as the referent group.

Counterfactual- the missing value that is estimated used by taking the differences in observed outcomes in causal analysis.

Effectiveness -The change in a dependent variable that is attributed to a specific cause or treatment.

Experiment- studies that investigate treatment effects of manipulable causes on specific outcomes to provide evidence for causality.

Hierarchical Linear Modeling (HLM) - commonly referred to as multilevel modeling, HLM is an analytic technique that is useful to examine data that are nested within one another such as students in classrooms, or teachers in schools.

Manipulable causes- Causes deliberately altered by the researcher when investigating the effectiveness of treatments.

Non-manipulable variables- variables that cannot be controlled by the researcher and are often included in as predictor or control variables.

Observational studies- studies that investigate causal relationships between non-manipulable causes such as demographic variables and outcomes.

Percent non-overlapping data (PND) - the percentage of data for which the estimated PS falls outside of the region of common support

Propensity score- the “conditional probability of assignment to a particular group, given a vector of covariates” (Rosenbaum & Rubin, 1983 p. 42)

Randomization- the process of using a random mechanism to assign subjects to treatment conditions. This process ensures that groups are balanced on all observed and unobserved covariates and any differences are random.

Root Mean Squared Error- the square root of the average sums of squares of the errors.

Treatment- the group receiving or exposed to a specific condition under investigation of its effectiveness.

CHAPTER TWO: LITERATURE REVIEW

This study investigated the behavior and performance of PS methods in multilevel settings. Accordingly, the review focuses on causal inferences using PS methods, and their appropriateness in multilevel studies. To provide a foundation the literature on causal inference, the theoretical framework for this study, is presented first. Included is a description of two distinct perspectives on the nature of causality. What follows is a discussion about the logic and use of PSs. Included here is a synthesis of current recommendations for applying PS methods within a single level context and a review of several empirical studies. Next, an overview to the theory and logic behind MLM is introduced. Lastly, research on PS in MLM will be discussed, empirical gaps will be identified, and a rationale for the proposed study will be offered.

Theoretical Framework

Originating as early as the 16th and 17th centuries, the concepts of causation and experimentation have influenced the development of Western Science in philosophy. Two distinct perspectives regarding the nature of causality and research are discussed within the literature: Rubin's Causal Model (RCM) and Campbell's validity framework. These two perspectives correspond to the two approaches to the estimation of causal effects in non-randomized studies (applied statistical methods and alternative research features), respectively.

RCM presents a framework for defining causal relationships and estimating treatment effects based on the counterfactual, sometimes referred to as potential outcomes (Holland, 1986). In contrast, Campbell's validity framework focuses on the inferences made from experiments and the potential threats to the validity of these inferences caused by various design related factors. Although often presented in isolation and operated independently rather than mutually (Shadish, 2010), these two perspectives are quite complimentary and share many common underlying

features (Rubin, 2010). RCM provides the foundation guiding the logic behind PSA while Campbell's commitment to improving causal inferences in non-randomized studies parallels the impetus of this study. Their unique perspectives provide a robust description of the theory of causality; thus, jointly serve as the theoretical framework for this study.

Rubin's Causal Model

RCM focuses on the precise mathematical and statistical properties related to causal inference, specifically the concept of potential outcomes (Rubin, 2010; West & Thoemmes, 2010). In RCM, treatment effects are determined by comparing the potential outcomes that would have been observed for an individual under different conditions. These outcomes are considered "potential" as each individual cannot be observed under various conditions simultaneously. In the simplest application of this model, there are two possible conditions (e.g. treatment and control) and each individual, i , has a potential outcome for each condition: $Y_i(0)$ for control and $Y_i(1)$ for treatment. For each individual, the treatment effect, τ_i , is defined as the difference between the two outcomes:

$$\tau_i = Y_i(1) - Y_i(0) \tag{2}$$

Given this definition, it becomes impossible to observe both outcomes for the same individual. For each individual in the experiment, one of the two outcome variables will be observed, while the other one will be missing. Subsequently, it is impossible to find the difference between the two outcomes for an individual, and the treatment effect for an individual can never be *observed*—the "Fundamental Problem of Causal Inference" (Holland, 1986, p. 947; Rubin, 1978).

RCM combats the fundamental problem using a statistical solution to *estimate* the ATE based on the expected value of the difference in outcomes, or a counterfactual model. Consider an experiment with two treatment levels, t (treatment) and c (control), where $Z=1$ when treatment is administered to an individual and $Z=0$ when the individual receives the control. Let Y represent the outcome variable of interest in this scenario. Each individual is assigned to one

condition, and therefore only one outcome is observed; the outcome for the condition the individual was assigned. The unobserved outcome, or missing outcome, is considered the counterfactual. Equation 3 represents the observed outcome, Y_i for an individual:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0) \quad (3)$$

In the counterfactual model, each individual will have one observed outcome Y , dependent upon Z , and the counterfactual will be missing. Observed outcomes for each condition can be averaged, as these averages from the sample also correspond to the population average. The ATE of the population, U can be estimated from sample, u outcomes. The two distributions of observed outcomes $Y_{z=1}$ and $Y_{z=0}$ are formed by separate individuals and these distributions represent hypothetical distributions of the population, had all individuals received treatment and control, respectively (Lunceford & Davidian, 2004). Consequently, the differences between the aggregated outcomes represent the ATE.

This solution to the fundamental problem allows causal inferences to be drawn using outcome measures observed from different individuals (Holland, 1986). In some situations, it is not the ATE that is of interest but rather the average treatment effect on the treated (ATT). For example, if students were able to elect to participate in an intensive dropout prevention program, the real interest here may not be the average effect, but rather the program's effect on those who participated. This is theoretically understood as the difference in outcomes with and without treatment for only those who were treated (Caliendo & Kopenig, 2008; Holland, 1986; Rubin, 1973a, 1973b). Since outcomes for all treatment conditions cannot be observed for all units, RCM operates under several key assumptions discussed below (Rosenbaum & Rubin, 1983, 1984; Rubin, 2010).

Strongly ignorable treatment assignment assumption. Causal analysis and counterfactual models rely heavily on the assumption of strongly ignorable treatment assignment, which when satisfied suggests alternative explanations have been accounted for and there is no hidden bias in treatment effects. This assumption refers to the mechanism or process used to

assign individuals to conditions and requires the assignment to condition be independent and not associated with the outcome or other factors. When satisfied, causal inferences can be drawn for population U using the average observed outcomes for all u in U exposed to t and c only (Holland, 1986).

When u units are randomly assigned to conditions it is assumed that the cause of assignment mechanism Z , is statistically independent from the outcomes Y_{z1} and Y_{z0} . This assumption is considered to be satisfied through the randomization process; however, when randomization is not employed satisfying this assumption becomes complex as differences in observed outcomes may be attributed to alternative variables related to the assignment mechanism. Accordingly, well executed RCTs are able to produce unbiased estimates of treatment effects and are, therefore, the preferred method when investigating causal relationships. The strongly ignorable treatment assignment assumption becomes a key factor regarding the quality of estimates in non-randomized studies. Ignorability of treatment assignment can be assumed if all the covariates that affect the treatment assignment have been accounted for, so that there are no unobserved covariates that may influence the estimates. If ignorability holds, one can obtain unbiased treatment effect estimates.

Stable unit treatment value assumption. Since outcomes for all treatment conditions cannot be observed for all units the outcomes from different units are compared. Therefore, experiments operate under the stable unit treatment value assumption (SUTVA), a strong independence assumption. More formally, SUTVA is defined as an "a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive" (Rubin, 1986, p. 961). Simply, SUTVA assumes the outcomes from two individuals, irrespective of treatment assignment, are independent from one another. When operating under SUTVA, statistical solutions can be applied to estimate the ATE over a population (Holland, 1986).

Campbell's Validity Framework

Non-randomized studies lack desired properties for causal inference and when causal inferences are incorrect or invalid, there are several potential reasons for this imprecision, referred to as "threats to validity" (Shadish et. al., 2002). Campbell's work focused on identifying the potential threats which often materialize in field settings and subsequently designing methods to account for them (Maxwell, 2010; Rubin, 2010; Shadish, 2010; West, & Thoemmes, 2010). A central focus of Campbell's framework was his distinction of the two inferences made through experimentation. The first one being whether the independent variable or manipulable cause can produce a significant effect (internal validity), and the second one being the identification of the different populations, settings, variables, and conditions to generalize the significant effect (external validity) (Campbell, 1957).

The term internal validity has been adopted in the social sciences and psychological literature with meanings that vary from Campbell's original definition (Shadish et al., 2002). In this context, internal validity refers to the extent to which the approximate truth is captured in an inference (Shadish et al., 2002). Validity can be considered a "property of the inferences" made, rather than a property of the design or methodology, because one design may yield more or less valid inferences in various circumstances (Shadish et. al 2002, p. 34). Cook and Campbell (1979) refined the distinction between internal and external validity and created a taxonomy of four different validity types to classify the different threats: (a) statistical conclusion validity, (b) internal validity, (c) construct validity and (d) external validity.

Statistical conclusion validity refers to the validity of the inferences about the covariation between treatment and outcome and answers the question regarding the strength, magnitude, and reliability of the covariation between the presumed cause and effect. Internal validity includes the threats associated to the validity of the inferences about whether the observed covariation between the treatment and the outcomes reflect a causal relationship from the treatment to the

outcome. Internal validity helps researchers understand whether the relationship was indeed causal or whether the outcome would have occurred regardless of exposure to the treatment. Construct validity is defined as the validity of the inferences about the higher order constructs representing sampling particulars such as the observed persons, the settings, and the cause and effect operations (Shadish et al., 2002). Construct validity allows researchers to identify the general constructs involved in the persons, settings, treatments and observations used in the experiment. Threats to construct validity identify a mismatch between the operations of the various particulars by the study and the constructs used to describe those operations. Lastly, external validity refers to whether the causal relationship holds over variation in persons, settings, treatment variables and measurement variables, or in other words the interaction effects. External validity specifically answers questions regarding the generalizability of the causal relationship over variation on the various sampling particulars.

The identification of specific potential threats within each of the four validity types has provided researchers with various sources that may be the culprit for the potential imprecision of the estimates generated from non-randomized studies. This process of ruling out threats to validity has been described as a “falsificationist enterprise” (Shadish et al., 2002, p. 41) where the purpose is to “reduce the number of plausible rival hypotheses” (Campbell & Stanley, 1963 p. 36) which would increase the “degree of confirmation” (p. 36) in the inferences. Arguably, Campbell's most significant contribution was his quest to identifying the threats and developing research features to include in the study design in order to logically rule out plausible explanations when conducting field based research (Rubin, 2010; Shadish, 2010; Shadish et al., 2002; West & Thoemmes, 2010) and provide a greater amount of trust for the causal inference in non-randomized experiments. Campbell and Stanley (1963) introduced the term quasi-experiment, a designation for such research design features.

Causal Inference in Non-randomized Studies

Various statistical methods and procedures have been applied to adjust the data from observational studies to estimate causal relationship. Traditional matching methods, such as simple mean matching, pair matching (Rubin, 1973a) and multivariate matching (Rubin, 1976, 1979) are statistical methods that aim to equate the groups. Matching methods can be used to estimate both causal effects as well as non-causal relationships (Stuart, 2010). Traditional matching of individuals on relevant variables is complex.

As the number of related variables used to match increases, the number of combinations for individual matches between groups also increases exponentially (Cochran, 1965). For example, if there are 10 dichotomous covariates, there are 1,024 combinations to match on. Thus, historically these methods allow for matching to be done on a limited number of covariates. Additionally, with traditional matching methods there is a potential to lose a lot of data. For example, in a study with 671 individuals assigned to treatment and 523 assigned to control; only 23 pairs were able to be matched on six categorical variables (Stuart, 2010). When few covariates are used to match, estimates of treatment effects may seem unbiased (Cochran, & Rubin, 1973). However, the likelihood that one or two variables can account for the total variance explained by the treatment assignment is slim, and differences in treatment effects may be attributed to systematic differences between groups on variables not included.

In 1983, Rosenbaum and Rubin introduced the PS, a major advancement in causal analysis, specifically due to its ability to balance groups using a set of covariates reduced to a single score virtually eliminating the challenges with traditional matching (Shadish & Steiner, 2010). The ability to include many variables increases the likelihood of satisfying the strong ignorability assumption. When the assumption of strongly ignorable treatment assignment holds PSs can yield unbiased treatment effects (Stuart, 2010; Shadish & Steiner, 2010).

The Logic of Propensity Scores

Rosenbaum and Rubin (1983) introduced the concept of using balancing scores to group treatment and control units in non-randomized studies. A balancing score is defined as the "function of observed covariates, such that the conditional distribution of these observed covariates is the same for treated and control units" (Rosenbaum & Rubin, 1983, p. 42). For example, balancing score, $b(x)$, is a function of the observed covariates, x (Rosenbaum & Rubin, 1983). Individual units assigned to different conditions with the same, $b(x)$, are assumed to be comparable, therefore using balancing scores may mimic the properties of a randomized experiment, where groups are considered to be probabilistically similar. In other words, balance implies that the variables x , are statistically independent of treatment assignment; therefore cannot be a confounding influence and unbiased estimates of the treatment effect are possible, assuming no other confounders exist. Rosenbaum and Rubin (1983) introduced the PS, as a specific and coarse type of balancing score that collapses a set of covariates into a single score to balance the groups.

Rosenbaum and Rubin (1983) proved that if the PSs are balanced between the two groups, then all the covariates used to estimate the PS are also balanced on average in large samples. Therefore, when conditioning on the balanced PS estimates, the covariates are statistically independent of treatment assignment, thereby mimicking the expected properties of randomized experiments potentially producing unbiased treatment effect estimates. A key assumption of PSs and their capability for producing unbiased estimates is that x , the vector of covariates, contains all variables that may potentially bias the treatment effect estimate. More simply, PSs rely heavily on the strongly ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983). The process of applying PS to estimate treatment effects requires multiple decision-based steps. These steps are described in detail below.

Covariate selection

The selection and inclusion of appropriate covariates is an integral component in PS. The quality of the estimated PS depends upon its ability to remove hidden bias which is a function of the covariates included in the model. PS relies heavily upon the assumption of ignorable treatment assignment; therefore, in order to satisfy the ignorable assumption all variables related to treatment and outcome need to be included (Rubin & Thomas, 1996; Stuart, 2010). There is no statistical test for this assumption; its satisfaction is inferred through substantive knowledge of the possible confounding variables in the applied context (Shadish & Steiner, 2010; Stuart, 2010; Steiner, Cook, Shadish, & Clark, 2010). Therefore, even if a large number of covariates are included, there is no way to tell whether a confounding variable has been excluded, or overlooked. Bias from unobserved covariates is known as hidden bias (Guo & Fraser, 2010; Rosenbaum, 1987, 2002). The researcher is confronted with the responsibility of selecting the appropriate variables to satisfy the strongly ignorable treatment assumption. If important covariates related to treatment assignment are omitted, then there is potential for the PS to be biased, thus the resulting treatment effect estimates will also be biased (Rubin, 1997; Shadish & Steiner, 2010). The exclusion of potential confounding variables impacts the treatment effect estimates and consequently threatens the validity of the inferences (Rosenbaum & Rubin, 1983; Shadish & Steiner, 2010; Steiner, Cook, & Shadish, 2011; Stuart, 2010).

Estimation Methods

There are two basic methods for estimating propensity scores: binomial regression models and statistical learning algorithms (Shadish & Steiner, 2010). Binomial regression models, such as logistic regression models or probit models, belong to the large family of models known as generalized linear models and are currently the most common method of PS estimation (Shadish & Steiner, 2010). Generalized linear modeling (GLM) is a technique for modeling data with non-normally distributed response variables, such as binomial distributions (Agresti, 1996; McCullough & Nelder, 1989; Nelder & Wedderburn, 1972; Quinn & Keough, 2001).

GLM's, specifically logistic models, consist of three components. First is a random component, which is the response variable and its probability distribution, second is the systematic component which is represented by the predictors or covariates used to specify the model, and third is the link function which is used to connect the random and systematic components (Agresti, 1996; Quinn & Keough, 2001). The most common link function for binary data and logistic regression is the logit link which transforms the predicted probabilities into logits or the natural log of the odds. Thus, in multiple logistic regression a set of k predictors for a binary response Y is modeled so that the logit of the probability, π , that $Y=1$ can be generalized using the following equation:

$$\text{logit}[\pi] = \log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n \quad (4)$$

Here, β_0 is the average PS for the sample or the average probability in logit units of receiving treatment across individuals assuming all the covariates are centered and normally distributed. The parameter β_1 refers to the effect of X_1 on the log odds that $Y=1$ controlling for the other X s (Agresti, 1996, Hosmer & Lemeshow, 2000). The logit link is considered favorable because the probabilities are modeled as a linear function of the covariates and the outcome, the natural log of the odds is a continuous, normally distributed variable (O'Connell & Rivet Amico, 2010). In the context of PS analysis, logistic regression models estimate the probability of being assigned to the treatment, conditional on the covariates specified.

With logistic regression models, least squares estimation for the model parameters are not optimal, because the variance of the binary outcome for each individual is not constant across all predictors, but rather depends on the probability for each predictor (Agresti, 1996). With the non-constant variance, maximum likelihood (ML) estimates can yield smaller standard errors than least squares estimates; therefore, ML estimation is used in logistic models (Agresti, 1996; O'Connell & Rivet Amico, 2010; Quinn & Keough, 2010).

Statistical learning algorithms also known as machine learning or ensemble methods refer to techniques such as classification or regression trees, boosting, bagging, or random forests (Shadish & Steiner, 2010). These methods try to iteratively minimize imbalance in covariates, while taking into account any nonlinear relationship between covariates and treatment assignment (McCaffrey, Ridgeway, & Morral, 2004). For example, a classification tree uses a recursive algorithm to estimate a function describing the relationship between a multivariate set of independent variables and a single dependent variable, for example treatment assignment. Using one variable as its basis, the tree fitting algorithm splits the dataset into two regions. The chosen split is the one that minimizes the prediction error. Within a region defined by its splits, the estimated function is equal to the sample mean of the outcome variable for all observations that based on the value of their covariate are included within the region. The regions are each subsequently partitioned and this process occurs iteratively until the number of allowable splits is reached (McCaffrey et al., 2004).

Since PSs are meant to create comparable groups, after PSs have been estimated, it is good practice to assess the two groups based on their estimated PSs. One method is to examine the distribution of the estimated PSs between the groups, known as the region of overlap or common support region (Shadish & Steiner, 2010; Stuart, 2010; Thoemmes & Kim, 2011). A broad region of common support allows for causal effect estimates to be based on the full range of PSs in the sample, whereas a small common support region restricts the effect estimates. Often individuals who have estimated PSs that fall outside the common support region are dropped from the analysis. These individuals are usually located in the tails of the distribution, with estimated PSs close to 1 and 0. If PSs yield insufficient overlap then it may be that the model to estimate the PSs needs to be adjusted.

Conditioning methods

Once PSs are estimated, the next step is to utilize, or condition on them to balance the treatment and control groups. There are four general conditioning techniques: matching,

stratification, covariance adjustment, and weighting. The first three techniques were introduced with the PS in 1983 by Rosenbaum and Rubin while the last technique was introduced a few years later (Rosenbaum, 1987). There are multiple decisions to be made within each conditioning method.

Matching. Propensity score matching involves the formation of new data sets which include only those individuals matched. Individuals from treatment and control groups who share a similar propensity score are matched to balance the groups (Guo & Fraser, 2010; Rosenbaum, & Rubin, 1983, 1985; Steiner & Cook, 2013; Thoemmes, & Kim, 2011). There are four major factors to consider when creating a matched data set: the matching algorithm, matching with or without replacement, number of control units and the distance criteria. Often these decisions are made collectively, rather than independently.

There are several matching algorithms to select from such as kernel, greedy, or optimal (Austin, 2011; Steiner & Cook, 2013; Gu & Rosenbaum, 1993; Hill, Weiss, & Zhai, 2011; Stuart, 2010). Kernel matching uses weighted averages of all cases in the control group to estimate counterfactual outcomes. The weight is calculated by the propensity score distance between a treatment case and all control cases. The closest control cases are given the greatest weight (Smith & Todd, 2005). Greedy matching considers the individual scores to match on based on the best control unit matched to each individual treatment unit (Gu & Rosenbaum, 1993). Optimal matching does not consider the best matches for individual treatment units, but rather matches are made to minimize the overall distance for all units in the matched sample (Thoemmes, & Kim, 2011). Matching with replacement allows controls matched to be placed back into the control group to be matched to other treatment units. Therefore, control units may be matched with multiple treatment units. Matching without replacement does not allow a control unit previously matched to a treatment unit to be considered for matching again (Austin, 2011).

Several different options regarding the number of control units (k) to match with each treatment are available and include: 1:1, k :1, or full matching. The number of matches for each of the treated units has some theoretically based implications regarding the precision and efficiency of the estimators (Steiner & Cook, 2013). When using a 1:1 matching strategy, one treatment unit is matched to the most similar control unit. Unmatched control cases are then discarded and there is potential for a loss of data when estimating treatment effects. The k :1 strategy allows one treatment unit to be matched to multiple control units, where the number of control units is a constant k for all treatment units. Full matching divides the total sample into non-overlapping subgroups where each subgroup has at least one treatment and one control unit in an attempt to minimize the distance between units (Hill et al., 2011).

There are generally two primary methods to determine closeness in propensity scores: nearest neighbor matching, and nearest neighbor matching with a specified caliper distance. Nearest neighbor matching simply matches the treatment and control units based on the nearest PS, while nearest neighbor matching with a specified caliper distance sets limits for what will be considered tolerable range around individual treatment unit's propensity score within which acceptable matches to a control may be made (Rosenbaum & Rubin, 1985). Nearest neighbor matching with a specified caliper distance includes identifying a distance metric, (d_{ij}) to determine the criteria for close matches for a unit i . The distance metric quantifies the dissimilarity between pairs of observations from alternative assignment conditions (Steiner & Cook, 2013). When ($d_{ij} = 0$) then the two units are identical on all observed covariates. When the distance metric is nonzero, then there is a difference between the two units on at least one of the covariates. The larger the difference the more covariates are not identical. Calipers are generally defined as standard deviation units on the original covariate (Steiner & Cook, 2013; Lingle, 2009; Rosenbaum & Rubin, 1985; Thoemmes, & Kim, 2011).

Stratification. Stratification, also referred to as subclassification, divides or classifies the sample into strata based on their estimated PS. Cochran (1968) found that stratifying a

sample based on one continuous variable into five subgroups or quintiles eliminated 90% of the bias due to that one confounding variable. Rosenbaum and Rubin (1984) extended Cochran's findings to PS. Similar to Cochran's findings, Rosenbaum and Rubin found that stratifying the data into quintiles based on the estimated PSs eliminates approximately 90% of the bias due to the observed covariates that are confounders when estimating a linear treatment effect (Rosenbaum & Rubin, 1984). Stratification yields a series of subsamples of individuals with estimated PSs from both treatment and control group. The individuals' PS estimates within a subsample are much closer in range than the full sample (Rosenbaum, & Rubin, 1984).

Stratification is similar to full matching in that a series of smaller subsamples are created; except with full matching the number of subsamples are automatically selected and assigned while in stratification the researcher manually stratifies the sample into subsamples, or strata. The effects of each stratum are pooled across the strata to estimate the ATE (Rosenbaum & Rubin, 1984). Austin, (2011) likens this procedure to that of a "meta-analysis of a set of quasi-RCT's" (p. 408) because within each stratum the effect of treatment on outcomes is estimated by comparing the outcomes of the treated and untreated subjects.

Covariance adjustment. Using this approach to estimate treatment effects, the outcome of interest is regressed on an indicator variable denoting the assignment status (e.g., Z) and the estimated PS (Austin, 2011). This is similar to an Analysis of Covariance (ANCOVA) model. ANCOVAs are regression models that allow for both continuous and categorical predictors to be modeled simultaneously and test whether certain factors have an effect on the outcome of interest after removing the variance for which covariates account. The regression model used would reflect the nature of the outcome variable; for a continuous outcome an ordinary least squares (OLS) regression model would likely be used, while a logistic regression model would be used for a dichotomous outcome. This technique allows for all the treatment and control units to be retained; thus, there is no subsequent loss of data after conditioning (Steiner & Cook, 2013).

Covariance adjustment using the PS is a conditioning method that differs in three important aspects from the other conditioning methods. First, conditioning and estimating the treatment effect occur as a single step, unlike all other conditioning techniques, where the treatment effects are not estimated within the conditioning process. Second covariate adjustment is the only conditioning method that uses a regression model relating the outcome to the treatment status and the PS. Lastly, there are several regression assumptions that apply when using covariate adjustment with PS. First a linear relationship is required between the PS and the outcome; second, the linear relationship should be modeled correctly when estimating the PSs; and third, the regression model should be free from violations to collinearity (Austin, 2011; Thoemmes & Kim, 2011).

Weighting. First introduced by Rosenbaum (1987), weighting on the PS is conducted to ensure samples are representative of the population of interest. Weighting is often used in survey research, test equating, and norming to draw inferences from non-representative samples to a population (Hahs-Vaughn & Onwuegbuzie, 2006; Morgan & Todd, 2008; Rosenbaum, 1987). Rosenbaum (1987) introduced weighting as a PS conditioning method as an extension of poststratification, a method of standardization, where samples are divided into strata or subclasses, and the means of the strata are reweighted using population frequencies. In PS analysis, individual units are weighted based on their estimated PSs.

Inverse propensity weighting or inverse probability of treatment weighting (IPTW) assigns an individual's weight w_i as the inverse of the probability of receiving the treatment that the subject actually received, $(1-Z)$. This process accounts for the misrepresentation of the sample to the population in that, individual units underrepresented in the treatment or control group are up-weighted while units overrepresented are down-weighted. Weights are defined as a function of the estimated PS, e_i and the assignment status, Z_i :

$$w_i = (Z_i / e_i) + [(1 - Z_i) / 1 - e_i] \quad (5)$$

Weighting of the individual units is dependent upon the inference of interest (Hirano & Imbens, 2002; Hirano, Imbens, & Ridder, 2003; Lunceford & Davidian, 2004; Morgan & Todd, 2008; Steiner & Cook, 2013). If the ATE is of interest, then the inverse-propensity weights for the treated Z_1 and for the untreated Z_0 can be factored out using Equation 6. The difference in the weighted means for the treatment and control groups is the estimate of the ATE:

$$\tau = \frac{\sum_{Z=1} W_i Y_i}{\sum_{Z=1} W_i} - \frac{\sum_{Z=0} W_i Y_i}{\sum_{Z=0} W_i} \quad (6)$$

When the ATT is needed, only those in the untreated or control group receive a weight. Units in the treatment group receive a weight of 1 and units in the control group are weighted based on the ratio of their estimated PS to the inverse of their PS: This is also referred to as weighting by the odds.

$$W_{z_0i} = e_{z_0i} / (1 - e_{z_0i}) \quad (7)$$

A single formula to weight both treatment and control units when estimating the ATT, similar to Equation 5 can be expressed as:

$$W_i = Z_i + (1 - Z_i)e_1 / (1 - e_1) \quad (8)$$

Using the weights based calculated in Equation 8, the ATT can be estimated similarly to the ATE using Equation 6.

Evaluating the Accuracy of the Propensity Score Model

The quality of an estimated PS score relies on two aspects: to include the important and relevant covariates in the model and to correctly specify the functional form of the covariates in the model (Guo & Fraser, 2010). Correctly specified PSs are able to successfully remove bias; whereas misspecified PSs do not capture all of the covariates' potential for removing bias (Steiner & Cook, 2013). Although there are no statistical tests to provide guidance on model specificity, two central properties of the PS model can be examined to assess the accuracy of the model: the

balance property and the common support region. The balance property refers to the idea that in order to mimic the properties of a randomized experiment, the estimated PSs and covariates should be balanced between the two groups. As described above, the common support region provides a good descriptive measure regarding model accuracy. If the region of common support is small, then it could mean the groups are not comparable, the model used to estimate the PSs was not properly specified, or the covariates measured do not adequately satisfy the strong ignorability assumption.

Assessing balance can be done by comparing the distributions of the covariates and the estimated PSs before and after conditioning using standardized mean differences, statistical tests, or graphical representations. The most common method to assess balance is to compare the standardized mean differences for each covariate and the estimated PS before and after conditioning between groups (Austin, 2011; Harder, Stuart, & Anthony, 2010; Rosenbaum & Rubin, 1985; Rubin, 2001; Stuart, 2010). Alternative methods to assess balance include inferential tests of mean differences such as factorial ANOVA (Rosenbaum & Rubin, 1984), cumulative density functions, visual analysis, or bivariate correlations (Steiner & Cook, 2013). Additionally, multiple measures may be triangulated to assess balance such as the standardized differences of means of the PS, the ratio of the variances of the PS in both groups, and for each covariate the ratio of the variance of the residuals, orthogonal to the propensity score in both groups (Rubin, 2001). Once the researcher is comfortable with the model, PS analysis officially ends and the researcher continues on to analyze the treatment effect estimates using the newly adjusted dataset.

To summarize, the PS is a single number balancing score estimated using a large number of variables. The PS is a relatively new statistical method that can assist in reducing bias of treatment effect estimates in non-randomized studies. The quality of PS estimates and the success of the approach to balance groups and reduce bias rely on several key assumptions,

especially the strongly ignorable treatment assignment assumption and the correct specification of the PS model.

Practical Concerns with Propensity Score Analysis

Since the introduction of the PS as a statistical method to account for plausible alternative explanations of treatment effects in non-randomized studies, researchers have empirically examined the behavior, performance, and efficiency of specific PS methods in order to identify best methodological practices during each PS step through a variety of means such as simulation methods (e.g. Austin, 2009; Gu & Rosenbaum, 1993), within-study comparisons (e.g. Steiner, Cook, Shadish & Clark, 2010), and analyses of existing databases (e.g. Stuart & Green, 2008). Unfortunately, consensus on what constitutes best practices does not yet exist. Therefore, current discussions and perspectives for each PS step are presented.

Covariate Selection

Causal analysis and counterfactual models rely heavily on the assumption of strongly ignorable treatment assignment. Accordingly the ability of the PS to produce accurate estimates of the treatment effect also relies on this assumption. Ignorability of treatment assignment can be assumed if all the covariates that affect the treatment assignment have been accounted for, so that there are no unobserved covariates that will affect the estimates. Therefore, the choice of variables can impact the overall performance of PS. This step is crucial as there are no tests to assess whether this assumption has been satisfied. One criticism of the literature is the lack of guidance as to which variables to include or exclude (Heckman & Navarro-Lozano, 2004).

Theoretically, the best practice would be to simply include all possible confounders; however, in practice researchers are generally unable to identify all potential confounders (Austin, 2011). Rubin (2007, 2008) recommends conducting a PS analysis using a prospective approach, where potential confounders are identified during the design phase allowing researchers the ability to use substantive theory to plan to measure all necessary covariates.

However, PS analysis is most commonly applied retrospectively, during the analysis phase, where only measured and available covariates can be considered.

Selecting covariates based on statistical model building approaches, (e.g. backwards, forwards, or stepwise regression), or *t* ratios of treatment group differences has been criticized (Kelcey, 2011; Rubin, 2008). The purpose of the PS is not to optimize an information criterion, or to maximize the correct prediction parsimoniously, (Rosenbaum & Rubin, 1984; Shadish & Steiner, 2010; Stuart, 2010); therefore, such statistical techniques will more than likely not include the important observed covariates needed to achieve balance and remove bias. These techniques focus on predicting the treatment and run the risk of excluding or removing potentially confounding variables due to a lack of power rather than a lack of balance which may ultimately fail to remove bias (Greenland, 2008).

Perhaps the most commonly applied method is to consider all available covariates. Including a large set of covariates would maximize the likelihood of satisfying the assumption, or conversely, minimize the chances of inadvertently omitting a potential observed confounder (Stuart, 2010). Often a large number of covariates are included in the PS estimation model. When comparing smokers to non-smokers, Rubin (2001) included 146 covariates in the PS model. Thoemmes and Kim (2011) reported as many as 238 covariates were used in substantive PS applications. The discussion on covariate selection extends beyond a number, to the quality of the covariates and the degree to their confoundedness.

A number of simulation studies have examined the impact of selecting different types of covariates on the performance of the PS. For example, one study examined the relationship between covariate selection and the overall performance of PSs by manipulating the magnitude of the association with Z and Y and the estimation models (Austin, Grootendorst, & Anderson, 2007). Using simulation methods, 1000 datasets of N=10,000 were generated with nine binary covariates with varying associations to assignment and outcome (see Table 3).

Table 3
Covariate relationship to assignment and outcome

	Strongly associated with assignment X_1	Moderately associated with assignment X_2	Not associated with assignment. X_3
Strongly associated with outcome			
Moderately associated with outcome	X_4	X_5	X_6
Not associated with outcome	X_7	X_8	X_9

Note: Table adapted from Austin, Grootendorst, & Anderson (2007)

To understand how variable selection affects the performance of PS estimates, 20 different PS models were considered. Unbiased estimates of the treatment effect were observed when the PSs were specified by the following models: the true propensity score model ($X_1, X_2, X_4, X_5, X_7, X_8$), the potential confounder model ($X_1, X_2, X_3, X_4, X_5, X_6$), the true confounder model (X_1, X_2, X_4, X_5), and the non-parsimonious model (all nine variables). However, the model with only true confounders yielded a larger matched sample and the lowest bias estimate and MSE, while the potential confounders model produced the lowest bias and MSE when stratification was used. Additionally, findings indicated when either all measured variables or all variables related to selection were entered in the PS model balance was achieved on these variables. When a covariate was not included in the PS estimation then balance on this covariate was not achieved. Therefore, if balance on all observed covariates is desired then PS estimates need to be specified with all the observed covariates (Austin, Grootendorst, & Anderson, 2007).

This study empirically supports including all covariates in the model; however other studies indicate that including unrelated variables, and/or instrumental variables that is, variables related to assignment only (Wooldridge, 2009), may amplify bias and increase noise to the PS as well as the correlation between the PS and the assignment mechanism (Brookhart, et al., 2006; Pearl, 2010). For example, Wooldridge (2009) proved that including instrumental variables may asymptotically lead to a greater bias, and when no bias exists to begin with, the overall precision is reduced. Pearl (2010) refers to instrumental variables as "bias-amplifying" variables. Pearl

extended the discussion of covariate selection to nonlinear models and asserts that inclusion of instrumental variables will not reduce selection-induced bias in either linear and nonlinear models and found that including these variables in nonlinear models may introduce new bias (Pearl, 2010).

In a pair of simulation studies, Brookhart and colleagues (2006) investigated the performance of various PSs to estimate exposure effects. The first simulation study examined whether specifying PSs using different types of covariates impacted the overall estimate of the exposure effect. The second study manipulated the strength of the confounder's relationship to outcome as well as exposure to examine whether the inclusion of a single confounder altered the bias and variance of the estimated exposure. Researchers considered three covariates: a true confounder, a variable related to outcome but not exposure, and a variable related to exposure but not outcome.

In the first study, seven different PSs were estimated, each corresponding to the different combinations of covariate specification (3 single covariate models, 3 double covariate models and 1 triple covariate model). The bias, variance, and MSE of the exposure estimate, from all seven possible combinations of covariate inclusion were compared to a crude log relative rate estimate (c-statistic). Results indicated PSs specified by the confounder variable and the variable related to the outcome performed the best. This model yielded unbiased estimates with the smallest variance for both study samples. Additionally, results demonstrated an increase in variance with no additional decrease in bias when including variables related to assignment but unrelated to outcome. The authors assert the following three conclusions: (a) including such a variable within the PS model adds noise to the estimate of the PS and increases the correlation between the estimated PSs and the assignment mechanism, (b) variables related to outcome are empirical confounders and regardless of their association to exposure should be included in the model, and (c) caution against including all available covariates in the model as the specification strategy (Brookhart, et al., 2006).

The second simulation study found consistent results regarding the relationship between covariates and the performance of PSs. Results demonstrated increasing the strength of the association between a variable and exposure increased the variability of the estimated exposure effect, independent of the variable's associated strength to outcome. Researchers conclude that if one wishes to minimize the MSE then in studies with a small N, it may be beneficial to omit a true confounder from the PS model, when the confounder is strongly associated to exposure and weakly associated to the outcome. This conclusion is strongly underscored by the small study characteristic because the researchers noted as the study size increases the variance of the estimator decreases at a rate of $1/n$, while the bias due to the omitted confounder remains (Brookhart, et al., 2006).

Logically, and empirically, the best combination of covariates would be a parsimonious model that included only confounders (Austin, Grootendorst, & Anderson, 2007; Steiner, Cook, Shadish, & Clark, 2010). However, the loss of precision does not seem to be as much of a concern as does the potential omission of a confounding variable (Austin, Grootendorst, & Anderson, 2007; Shadish, Clark, & Steiner, 2008; Steiner et al., 2010; Stuart, 2010). One method to avoid is to rely solely on demographic variables to estimate the PS (Shadish et al., 2008; Shadish & Steiner, 2010; Steiner et al., 2010; Stuart, 2010; Thoemmes & Kim, 2011). Using a doubly randomized preference trial, a within-study analysis found that the PSs specified by demographic covariates only performed poorly, even worse than the estimates from the unadjusted quasi-experiment, as measured by the absolute bias, percent bias reduction, and MSE (Shadish et al., 2008).

Research has identified the importance and substantial impact on the PS regarding covariate selection, specifically, considering the strength of relationship with assignment and outcome. Effectively including covariates associated to the observed outcomes requires either a solid foundation of the theory related to the outcome, or the use of the observed outcomes when considering which covariates to use to estimate the PSs (Kelcey, 2011). However, Rubin (2008)

asserts the potential benefits of using observed outcomes to construct PSs are negated by the various consequences capable of diminishing the quality of the inferences, such as omitting key variables that influence treatment assignment. Recently, Kelcey (2011) examined the use of outcome proxies as a method to systematically consider covariates related to the outcome, without compromising the efficiency of the effect estimates by including a census of covariates or the diminishing the quality of the inferences by using the observed outcomes.

A set of Monte Carlo simulation studies was conducted to examine the extent to which observed outcome proxies and cross validation methods to approximate covariate's relationship with potential outcomes were similar to those using the observed outcomes and how the inclusions of different types of covariates in a PS model affected the treatment effect estimator and the covariate balance under different conditions. Four types of covariates with differing magnitudes were considered: covariates related to both outcome and assignment, covariates related to outcome but not assignment, covariates related to assignment not outcome, and covariates unrelated to outcome and assignment.

Several covariate combinations were considered when constructing the PS models: predictors of proxy only, union of predictors of the proxy and treatment, intersection of predictors of the proxy and treatment, predictors of treatment only, union of predictors of the outcome and treatment and using cross validation, intersection of predictors of the outcome and treatment using cross validation, and all possible covariates (Kelcey, 2011). Covariates were deemed proxy, treatment, or outcomes based on their ability to predict proxy, treatment, or outcome measured by information criteria such as Akaike Information Criteria (AIC) and Bayesian information criteria (BIC) garnered through stepwise procedures. To determine the effectiveness of estimating covariates' relations to the potential outcomes using an outcome proxy, the outcome proxies and cross validation estimates were compared to those using the observed covariate associations. Specifically, a pretreatment outcome proxy correlated to the true outcome at the 0.7, 0.5, and 0.3 levels and a 50% cross validated subsample were considered (Kelcey, 2011).

Findings indicated PS methods within a correlation condition tended to yield comparable results, however this was not the case across correlation conditions. Consistent with the previous studies, the MSE and the bias were lowest when covariates strongly related to the outcome were included— either related to treatment and outcome or related to the outcome only. Specifically, PSs using only the important predictors of the observed outcomes yielded similar MSE between the proxy approach and the observed outcome approach. This held true with small proxy outcome correlations and small sample size ($n=500$). Conversely, the cross validation approach produced a higher MSE in the small sample size and a comparable MSE in the large sample ($n=2500$). However, this relationship was not the same when examining the bias estimates.

The cross validated approach yielded bias estimates comparable to those of the observed outcomes while the proxy approach generated higher levels of bias. Similar findings were noted when using the intersection of covariates related to treatment and outcome were considered. When the union of the covariates predicting treatment and assignment were used in the model, the proxy approach estimates, both MSE and bias, were very similar to the observed outcomes, especially when the sample size was large or when the proxy-outcome correlations were high (Kelcey, 2011).

When comparing the selection methods across conditions results were not as homogenous. Methods that included predictors of outcome proxy only or in union or intersection with the predictors of treatment produced lower MSE and comparable bias than the methods considering only the covariate treatment associations. Conversely, methods using the cross validated observed outcomes tended to produce lower bias estimates with the MSE dependent upon how the treatment predictors were combined (Kelcey, 2011). In terms of covariate balance, differences among the approaches were quite clear. Approaches focusing only on the covariates predicting the observed outcomes, or those including the intersection of the observed outcome and treatment predictors (true confounders), demonstrated large imbalances on the predictors related to assignment. Specifically, when more variables were specified, models tended to

provide a coarser balance on many covariates, while parsimonious models yielded a more precise balance on selected covariates only (Kelcey, 2011).

Currently, based on the literature the best method to use would be to include all available and measured covariates and to avoid relying solely on demographic variables to remove bias. The assumption of strong ignorability is extremely crucial to the reduction in bias estimates. Researchers found including poorly measured confounders result in better performing PSs than PS estimates that were specified with perfectly measured covariates but omitted one confounding variable (Steiner, Cook, & Shadish, 2011). Therefore, researchers should include covariates that allow them to confidently assume strong ignorability.

Estimation Methods

Currently, almost exclusively, binomial regression models, more specifically, parametric linear logistic regression with observed covariates as predictors for a binary treatment assignment (Luellen, Shadish, & Clark, 2005) are used to estimate the PS (Shadish & Steiner, 2010). One of the criticisms' for using logistic regression models to estimate PS scores is their sensitivity when the relationship between the covariates and the assignment function is nonlinear (Shadish & Steiner, 2010). In contrast, statistical algorithmic approaches automatically consider and account for nonlinear terms within the estimation model (Luellen, et al., 2005; McCaffery, Ridgeway, & Morral, 2004; Shadish & Steiner, 2010; Thoemmes & Kim, 2011), yet are rarely used.

Shadish and Steiner (2010) offer four reasons as to why binomial regression techniques are favored. First, PSs can easily be estimated by researchers using logistic regression, while understanding the basis of the learning algorithms is more complex and challenging (Shadish & Steiner, 2010). Second, since statistical learning algorithms are rarely used little empirical evidence on their relative ability to eliminate bias exists. Third, when the initial PS estimates do not yield balance for groups based on the covariates, using logistic regression researchers can easily adjust and re-specify the model. In contrast, with statistical learning algorithms, it is

unknown how to recalibrate the algorithmic procedures to garner better balance. Lastly, statistical learning algorithms may seem advantageous with regard to identifying the best prediction of assignment membership; however, these methods tend to favor the best fit or the information criteria of the estimation method rather than focusing on achieving balance on the covariates. Consequently, these methods may produce less optimal estimates of the PS (Shadish & Steiner, 2010).

Understanding the linear nature of logistic regression modeling, Dehejia and Wahba (1999, 2003) were one of the first to specify models with higher order terms such as polynomials and quadratic interactions. Results demonstrated estimates generated from models regressing a more non-linear function, were less biased than their linear counterparts. Researchers assert the importance of specifying models sufficiently in order to maximize the PSs' ability to remove bias (Dehejia & Wahba, 1999).

A handful of studies have compared the effectiveness of different statistical learning algorithms to logistic regression techniques. These studies indicate PSs generally perform better when estimated using various statistical algorithmic procedures over logistic regression; however upon critical analysis results tend to prompt more speculative questions rather than offer definitive solutions.

For example, Setoguchi and associates (2008) investigated the performance of various PS estimation methods for estimating exposure effects through simulation methods. Researchers considered seven different scenarios for data generation crossing various combinations corresponding to the degree of linearity and additively for the associations between the exposure and covariates. For each of seven scenarios, PSs were generated using four different estimation methods. Specifically, logistic regression models, classification trees with and without pruning and neural network methods were examined. Additionally researchers examined whether model *c* statistics predict bias and efficiency of the exposure estimates in the outcome model. Results indicated the PS models created using neural networks yielded the least biased estimates for many

scenarios, while the logistic regression models were robust to model misspecifications. However, other than stating "neural networks with 1 layer and 10 hidden nodes" (Setoguchi et al., 2008, p. 3) no other description of this method or citation was provided, rendering the replicability of this study difficult.

Lee and associates (2010) examined the performance of various PS estimation methods when conditioning using PS weights. The researchers simulated data similar to Setoguchi and colleagues (2008) only slightly modifying the structure. A binary exposure A with the exposure probability at the average of covariates was ≈ 0.5 , a continuous outcome Y with a population exposure effect $\gamma = -0.4$, and 10, covariates, 5 continuous and 5 binary with varying associations to exposure and outcome as well as between variables were modeled (see Table 4).

Table 4
Data structure for Lee and associates (2010) simulation study

		Type of Covariate	Between variable correlation
	W_1	True confounder	$r_{pbW_5} = 0.2$
	W_3	True confounder	$r_{\Phi W_8} = 0.2$
Binary	W_6	Exposure predictor	$r_{pbW_2} = 0.9$
	W_8	Outcome predictor	$r_{\Phi W_3} = 0.2$
	W_9	Outcome predictor	$r_{pbW_4} = 0.9$
Continuous	W_2	True confounder	$r_{pbW_6} = 0.9$
	W_4	True confounder	$r_{pbW_9} = 0.9$
	W_5	Exposure predictor	$r_{pbW_1} = 0.2$
	W_7	Exposure predictor	--
	W_{10}	Outcome predictor	--

Samples of $n=500$, $n=1000$, and $n=2000$ were replicated 1000 times for seven different scenarios that varied the degrees of linearity and additivity, specified with quadratic and interaction terms (Lee, Lessler, & Stuart, 2010). Specific properties for each of the scenarios were described as: additive and linear (main effects only), mildly non-linear (one quadratic term), moderately non-linear (three quadratic terms), mildly non-additive (three two-way interaction terms), mildly non-

additive and non-linear(three two-way interaction terms and one quadratic term, moderately non additive (10 two-way interaction terms) and moderately non-additive and non-linear (10 two-way interaction terms and three quadratic terms) (Lee et al., 2010, p. 339). PSs for each of the seven scenarios were generated using six different estimation procedures: logistic regression with a main effect for each covariate, basic classification tree, pruned classification tree with a cost-complexity parameter, bagged classification tree with 100 bootstrap replicates, random forests, and boosted regression trees (Lee et al., 2010).

Results indicated that as sample size increased the covariate balance increased under all scenarios, thus resulting in less biased effect estimates for all estimation methods. Consequently, the logistic regression, classification tree, and pruned classification tree methods produced poor confidence interval coverage when $n=2000$ as error terms shrink with larger sample sizes yielding more precise estimates. Overall the logistic regression model consistently and adequately balanced covariates with main effect terms; however when the models did not account for interactions and nonlinearities, the estimates were substantially biased. Alternatively, bagging, random forests, and boosted methods demonstrated consistently favorable estimates regardless of sample size or the extent of non-additivity or non-linearity (Lee et al., 2010). Additional studies comparing estimation methods also tended to favor algorithmic procedures over logistical regression models (e.g. Luellen et al, 2005; McCaffery et al, 2004). However, while findings indicated algorithmic methods performed better, the process by which the logistical regression models were specified were questionable and conclusions may be biased.

One study compared the treatment effects of a drug rehabilitation program using a generalized boosted model (GBM) and two logistic regression models (McCaffery et al, 2004). Statistical significant tests were used to select covariates for the logistic regression models. The first model included covariates with significant ($p<.05$) bivariate relationships with assignment and the second one with a more relaxed association ($p<.20$). Overall, the GBM model yielded smaller prediction errors, smaller absolute effect sizes, and balance groups compared with the

logistic regression models (McCaffery et al., 2004). All 41 pretreatment variables were automatically included by the GBM procedure. However, the number of statistically significant covariates was not provided for the logistic regression models; therefore it is assumed that all three estimation methods included a different number and combination of covariates. Since statistical algorithmic methods automatically select the covariates and model the functional form of the estimation procedure it can be assumed that the significance test may have omitted a potential confounder variable. Thus, if a potential confounder was omitted in either logistic regression model, especially one with a high association to outcome, the model would not perform as well as one where the covariate was included. It would be interesting to compare the relative performance when specifying logistic regression or probit models similar to automatic specification of algorithmic procedures.

Similarly, Luellen and associates (2005) used secondary data to compare the relative effectiveness of estimating PS using logistic regression, classification trees, and bagging bootstrap replicates. These estimates were compared to estimates from the randomized experiment. To estimate the logistic regression models, a backward stepwise logistic regression approach was used. All 25 covariates were included and researchers retained those covariates that significantly predicted group membership at $p < .50$. For participants with missing data, the pattern of the missing data was consistent therefore only 8 of the 25 covariates were retained in the model. Conversely, two classification tree models, and three bootstrap replicates all used the complete set of 25 covariates to estimate the PSs. Additionally, using classification trees, balance was not obtained according to the criteria set by the researchers who noted "we are aware of no other advice in the literature about how to refine the classification tree model to further obtain balance" (p. 542). Findings were inconsistent across PS models and conditioning strategies for the two different treatment effect outcomes being estimated (Luellen et al. 2005). Given the inconsistent findings researchers caution against using PS as a method altogether; however, the

argument could be made that findings were as inconsistent as the comparability among the models.

Preliminary research indicates algorithmic estimation methods may perform better than logistic regression methods when the functional form for the model is not linear. Conversely, logistic regression models can be easily adjusted and refined if balance is not achieved. However, logistic regression methods are sensitive to assumption of selection on observed covariates (Dehejia & Sadek, 1999), specifically when the association between assignment and covariates are nonlinear. When logistic regression models are specified correctly, the differences in methods are marginal and are overshadowed by the ability to control and adjust the logistic regression models as opposed to the black box nature of algorithmic methods. Currently, the empirical evidence supports using logistic regression models and to consider including selected interaction and polynomials terms within the model when using PSs to adjust for group differences in applied studies. In addition, there is a need for researchers to continually investigate the relative performance of various estimation methods and models.

Conditioning Methods

Theoretically, if the assumption of strongly ignorable treatment assignment is satisfied, conditioning by matching, stratification or covariance adjustment would all produce unbiased estimates of the treatment effects (Rosenbaum & Rubin, 1983). However, given the impossibility of empirically testing this assumption, different conditioning methods have produced different estimates in practice. Shadish and Steiner (2010) criticize the lack of empirical work comparing the relative effectiveness of the different PS conditioning methods to be able to confidently promote one conditioning method over another for various data conditions.

Matching, as a conditioning strategy has received a substantial amount of attention. As evidenced by Thoemmes and Kim's (2011) and Austin's (2008a) review of PS applications, matching is the most common and widely used conditioning method. In fact, Austin's (2008a) review focused solely on the variants of matching applied in the medical literature. Consequently

matching is by far the most examined conditioning strategy specifically the different factors to matching (e.g. algorithm, number of control units, replacement, and criteria).

Early research by Gu and Rosenbaum (1993) found that optimal matching generally outperforms any greedy matching algorithm. Greedy matching pairs each treatment with the nearest control; however, the best possible match may not be made across units. Optimal matching uses an algorithm to iteratively find a solution which will minimize the average PS distance across all pairs to obtain the highest degree of balance for the sample.

When considering how many control units to match, Imbens (2004) suggests "using only a single match leads to the most credible inference with the least bias, at most sacrificing some precision" (p.14). Ming and Rosenbaum (2001) reported that when using multiple control matches, selecting up to 3 matches can decrease bias, but little efficiency is gained when selecting more than 5 matches. Austin (2010) found similar results which indicated an increase in the number of untreated subjects matched to treated subjects increased the bias of the treatment effect as well as the overall precision and recommends matching be done with 1 or 2 untreated subjects. Using 1 untreated subject will minimize bias, where using 2 may result in improved precision without impacting the bias greatly (Austin, 2010). Additionally, Ming and Rosenbaum (2001) found that optimal matching with multiple controls cannot be obtained by appending the best available matches to an optimal pair matching. In contrast, 1:1 matching results in a loss of data, while full matching allows multiple matches for both treatment and control group members, which allows for the most possible number of cases retained in the sample. When using $k:1$, the efficiency can be increased, but the precision is decreased because with the increase in the number of matches less similar cases are matched (Steiner & Cook, 2013).

Matching with replacement has the potential to decrease bias because controls chosen are based on the individual unit's most similar match. In contrast, matching without replacement matches individual units to the most similar match from the pool of those not already matched. One drawback to matching with replacement is that treatment effects may be potentially

estimated using a small select group of control units, subsequently distorting the generalizability of the inferences (Stuart, 2010).

In an effort to avoid making poor matches, restrictions or criteria can be imposed on the distance measures. Austin (2009) investigated the relative performance of several matching distance measures: matching using calipers of 0.2 and 0.6 of the standard deviation of the logit of the propensity score, matching on the PS using calipers of 0.005, 0.01, 0.02, 0.03, and 0.01, and finally 5 to 1 digit matching where the treated subjects are matched to untreated subjects on the first five digits of the PS. Results indicated that using calipers of 0.6 of the standard deviation of the logit of the propensity score yielded the greatest percentage of matched pairs, while the lowest percentage of matched pairs were found in models using calipers of 0.005, both regardless of prevalence of exposure. Balance measures were all comparable across prevalence of exposures and methods. The relative bias was the largest when using calipers of 0.6 of the standard deviation of the logit of the propensity score, and increased as the prevalence of exposure increased. Using a caliper of 0.2 of the standard deviation of the logit of the propensity score resulted in less bias, but the lowest bias was observed when calipers of 0.05, 0.01, 0.02, and 0.03 were used (Austin, 2009). Additionally, some advise combining conditioning methods such as regression adjustment with matching (Rubin & Thomas, 2000).

Although matching is the most frequently used conditioning method, other methods also have both favorable and less than desirable qualities. For example, with stratification, covariance adjustment, or weighting, exact matches do not need to be estimated. Additionally, for covariance adjustment and weighting the full sample may be retained. Several studies have recently reported the results when using multiple conditioning methods. Findings indicate some conditioning methods perform better in reducing the bias, while other methods yield smaller error variances in the treatment effect estimates, increased statistical power, and smaller confidence intervals (Austin & Mamdani, 2005). Two additional studies found, on average, the different conditioning methods produced comparable results (Cook & Steiner, 2010; Steiner et al., 2010).

In contrast, some studies revealed differences in estimates of the treatment effect among the conditioning methods (Harder, Stuart, & Anthony, 2010; Kurth, Walker, Glynn, Chan, Gaziano, Berger, Robins, 2005; Lunceford & Davidian, 2005).

When comparing the performance of stratification to weighting with simulated data, where the assignment was generated using a Bernoulli distribution such that lower response on the covariates indicated a higher probability to be treated, weighting consistently produced unbiased estimates while stratification estimates were inconsistent (Lunceford & Davidian, 2005). Kurth and associates (2005) found including units with a very small probability impacted overall results and results were more trustworthy when those units were removed. This would align with the original intent to stratify the sample with the units who fell within the common support region; however in practice many studies stratify across the entire PS range which may be the reason why stratification seems to yield unbiased estimated in some studies. Lastly, Harder and associates (2010) did not find one conditioning strategy to outperform others, but found the interaction of the estimation method, model specification, and conditioning strategy seemed to matter.

Evaluating the Accuracy of the PS models

Accurately specifying PS models by including relevant covariates to satisfy the strongly ignorable treatment assumption and modeling their form properly is important to the quality of the PS estimates (Guo & Fraser, 2010). Unfortunately, there are no statistical tests to prove whether the ignorable assignment assumption has been satisfied. Rubin (1997) recommends conducting a sensitivity analysis to test whether the model is sensitive to potential violations of this assumption. A sensitivity analysis does not check to see whether you have correctly specified the model and selected the correct covariates, instead it is concerned with the "bias that results from not observing all the relevant covariates" (Imbens, 2003, p 126.). More specifically, this is testing if a potential unobserved confounder was omitted whether the overall estimates

would be impacted. However, currently the literature is not clear as to how to efficiently and confidently conduct a sensitivity analysis.

Currently PS models are evaluated against their ability to achieve balance. Few studies have examined the different methods for assessing model accuracy or balance. Consequently, applied studies often do not report balance estimates (Austin, 2008a; Thoemmes & Kim, 2011). Rosenbaum and Rubin (1984) suggested using a factorial ANOVA to assess balance. However, in his review of the applied matching methods Austin (2008a) asserts a strong criticism against using significance tests to assess the balance. He argues that significance testing is meant to test a null hypothesis against the population and whether the sample represents the population is irrelevant in balance testing. In addition, the reduction in sample size with matched data, may impact the significance of the imbalances even if they remain the same in absolute terms. For example, Imai, King, and Stuart (2008) randomly discarded control individuals which seemed to lead to an increase in balance, but this increase was masked simply because of a reduction in power.

Furthermore, studies have assessed whether different statistical tests were helpful in adequately evaluating model accuracy and balance and found that both the goodness of fit and c -statistic were not able to consistently identify models where a confounder variable was deliberately omitted (e.g. Brookhart et al., 2006; Weitzen et al., 2005). Austin (2008a) chastised using statistical significance tests as these tests are used to generalize to a population and balance is sample dependent. Currently, the most common and accepted method for assessing balance is to estimate the standardized difference in means for the groups on the covariates (Shadish & Steiner, 2010; Stuart, 2010).

Cohen's d is the most common measure of the standardized difference in means and is interpreted similar to an effect size. Equation 9 denotes the standardized difference in means d of a continuous covariate, where \bar{X} is the mean and s_i^2 is the variance for a particular group i and

averaging the group variances prior to taking the square root results in a pooled standard deviation (Stuart, 2008). This method assumes the groups are equal size.

$$d = (\bar{X}_{treatment} - \bar{X}_{control}) / \left[\sqrt{(s_{treatment}^2 + s_{control}^2) / 2} \right] \quad (9)$$

To calculate the standardized difference in means for a dichotomous variable, Austin (2009) suggests taking the difference in proportions, p between treatment and control groups will be comparable to the standardized difference for a continuous group.

$$d = (p_{treatment} - p_{control}) / \sqrt{\left[p_{treatment}(1 - p_{treatment}) + p_{control}(1 - p_{control}) \right] / 2} \quad (10)$$

Finally, currently, there are no rules of thumb regarding what constitutes balance between the groups. The smaller the standardized difference the better, but at what point is a difference considered imbalanced is not yet clear (Ho, Imai, King, Stuart, 2007; Sekhon, 2007). Some suggest $d < 0.20$ or 0.25 , while others recommend achieving as close to zero as possible (Shadish & Steiner, 2010).

Regardless, the ability of PS estimates to reduce bias is dependent upon whether or not the PS estimates are able to adequately balance the groups, in other words, mimic the properties of a randomized experiment, where group differences are a result of sampling error. Some posit that if balance is not achieved, and there is a lack of overlap between treatment and control group, then the sample may not be adequate for removing bias and estimating treatment effects (Shadish & Steiner, 2010; Stuart, 2010).

The Overall Effectiveness of Propensity Score Methods

Much of the literature on causal inference discusses the quality and behavior of various statistical and research design methods used to estimate causal relationships. Specifically within the PS literature, research compares the performance of PS methods against alternatives such as regression adjustment models, or regression discontinuity designs. Shadish and Steiner (2010)

claim that currently there is not enough conclusive evidence to suggest PS methods work better than alternative ones.

There has been quite a bit of discussion within the field of econometrics regarding the ability of propensity score methods to replicate experimental results. Lalonde (1986) examined the extent to which nonexperimental estimators could replicate the unbiased experimental estimate of the treatment impact when applied to a composite dataset of experimental treatment and nonexperimental comparison units. His seminal study estimated the impact of the National Supported Work (NSW) Demonstration on post intervention income levels. He found that nonexperimental estimators were not able to produce accurate estimates relative to the experimental benchmarks and were sensitive to the specification (Lalonde, 1986).

Dehejia and Wahba (1999) used Lalonde's (1986) data to evaluate whether PS methods are able to yield unbiased estimates of the treatment impact by comparing the results with his experimental and nonexperimental treatment effect estimates. Their comparative analysis indicated that while PS methods are not applicable in every setting, when the range of PS estimates overlap between treatment and control groups, estimates of the treatment effect may be comparable to those estimated under experimental conditions. However, these findings were not accepted without criticism. Smith and Todd (2001, 2005) asserted that alternative econometric methods perform better than PS methods and Dehejia and Wahba's (1999) findings are heavily affected by the exclusion of over 40% of the Lalonde's (1986) original data. In addition, Smith and Todd (2001, 2005) extended the PS methods and specifications to additional samples to compare the overall resulting bias and found PS methods were unable to replicate desirable findings across samples.

In 2005, Dehejia replied to Smith and Todd's (2001, 20015) critique of their findings by asserting two major points. The first point being that Dehejia and Wahba (1999) never insinuated PS methods were optimal or better than other methods, rather their work emphasized the conditions for which estimates may be made and discuss the methods ability to evaluate the

quality of comparison groups. Second, Dehijia (2005) posits that PS methods, specifically the specification models (i.e. the covariates selected and the functional form modeled) to estimate the PSs are sample dependent and are not meant to generalize to a population; therefore, the conclusions made by Smith and Todd (2001, 2005) are expected. Motivated to further investigate PS methods using Lalonde's (1986) data set, Michaelopolous, Bloom and Hill (2004), compared treatment effect estimates when controls were taken from different sites. This analysis found that PS methods work less well when comparing treatment and control groups from different social contexts or settings. Similarly, in their within study experiment, Shadish and others (2008) found that when controlling for all the covariates in a regression equation without estimating a PS, effect estimates were just as effective in removing bias as were PS methods. However, some studies have found that alternative methods behave in a more optimal manner than PS methods for certain treatment effect outcome measures.

In 2005, researchers conducted a systematic review of studies that used both PSA and traditional regression analysis of their observational data to examine whether different methods gave different results when adjusting for confounder bias (Shah, Laupacis, Hux, & Austin, 2005). A total of 43 studies were included and from these studies 78 assignment-outcome associations were found and odds or hazard ratios for each association under both methods were compared. Only eight of the 78 associations differed statistically between the two methods where the traditional regression method yielded statistical significance with the exposure-outcome association while the PS methods did not. Authors conclude that each method produced similar results when trying to adjust for the confounding and while PS methods produced slightly weaker associations, it was noted that some of the studies included did not incorporate PS methods adequately (Shah et al., 2005). A similar review of published studies found results comparable to Shah and associates (2005); however, the perspective offered regarding these findings was quite the opposite (Sturmer, Joshi, Glynn, Avorn, Rothman, & Schneeweiss, 2006). This study found

that while the applications of PS methods were increasing, the estimates were not substantially different than traditional methods and questioned the increased use of this method.

Several studies have compared the performance of PS methods with different outcome measures (Austin 2007, 2008b; Austin, Grootendorst, Normand, & Anderson, 2007). Results varied, indicating in some instances certain PS methods are not appropriate to measure certain outcome measures, and with other outcome measures, optimal results were produced with certain PS conditioning techniques. For example, in one study, researchers considered how the four different PS conditioning techniques behaved when estimating conditional odds ratio, hazard ratio, and rate ratios (Austin, Grootendorst, Normand, & Anderson, 2007). Results suggest PSs may not be the best statistical method to use when the treatment effects are measured in odds ratios or hazard ratios estimates. Specifically, researchers found when there was a true non-null treatment effect for either of these outcome measures, then PS methods produces biased estimates of the true conditional treatment (Austin, Grootendorst, Normand, & Anderson, 2007). For these measures, regression adjustment models yielded unbiased estimates of the treatment effect. Conversely, when rate ratio was the outcome measure for count data, PS methods did not introduce bias; however, neither did regression adjustment methods. Additional research indicates special considerations need to be taken when estimating marginal odds (Austin, 2007) and relative risks (Austin, 2008b)

Overall, the empirical research on PSs can be summarized into three concluding points. First, a great deal of research has been dedicated to examining the different conditions for which various PS methods work well. Second, paradoxically, the plethora of research on PS methods has provided many inconclusive findings. Lastly, there continues to be a need to empirically investigate PS methods in order to confidently accept or reject this method when identifying causal relationships from non-randomized studies or observational data.

Rubin (1997) stated that PSs cannot adjust for unobserved covariates, work better with larger samples, and do not behave in the same manner to a covariate associated to assignment but

not to outcome as it would to a covariate with the same relation to assignment but heavily associated to the outcome. Results from the empirical studies above not only support these three caveats but also give way to the emergence of additional questions regarding the strengths and limitations to PS analysis. Much of the empirical research presented thus far aimed to identify effective methods or best practices to apply PSs as a method for adjusting non-experimental data when drawing casual inferences. However, currently these goals are pending continual research.

Multilevel Modeling

Thus far, the discussion of causal inference and the PS has been limited to single-level settings. However many settings are multilevel, especially those where causal inferences are often drawn from non-randomized experiments or observational studies. Analyses that ignore the nested structure of data can result in misleading or inaccurate results because the assumption of independence is violated. Often in educational settings, certain schools, demographic areas, or neighborhoods have differing student achievement levels, and arguably certain settings are predisposed to offer advantageous learning environments over others (Oakes, 2004). Additionally, assignment to condition may be a result of the contextual factors of the cluster. This is where causal inference is complicated as each single unit potential outcome is not only dependent on treatment assignment but also on cluster membership and any cluster level contextual factors—a violation of SUTVA (Thoemmes, 2009).

MLM is a family of statistical analyses used to evaluate nested data (Raudenbush & Bryk, 2002). MLM models improve the estimation of individual effects in nested data by accounting for the dependencies among the units, adjusting the standard error properly, and partitioning the variance at all levels (Raudenbush & Bryk, 2002). Additionally, MLM allows for cross-level interactions, which explain how variables measured at one level affect the associations occurring at another level (Raudenbush & Bryk, 2002). In other words, using MLM with nested data allows researchers to investigate how much variability in an outcome is associated to treatment and control group differences for individuals both within and between clusters and the

extent to which different various within and between group factors account for the variability (Ferron, Hogarty, Dedrick, Hess, Niles, & Kromrey, 2008).

Consider an example where individual students are nested in classrooms and the effect of a new teaching program on student achievement is being evaluated. Equation 11 represents the relationship between student achievement, Y , for an individual in classroom j and the average student performance, β_0 , for classroom j .

$$Y_i = \beta_{0j} + e_{ij} \quad (11)$$

Here, β_{0j} is the average student achievement for each classroom with σ^2 as the variance of e_{ij} , the error term. A level-2 model can represent how the β_{0j} varies across classrooms, or whether the average student achievement is different among the classrooms:

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \quad (12)$$

In this equation, γ_{00} is the average of all the classroom averages and τ_{0j} is the variance of the μ_{0j} . A combined model can be formed by combining the level 1 and level 2 models.

$$Y_{ij} = \gamma_{00} + \mu_{0j} + e_{0j} \quad (13)$$

In MLM, both level 1 and level 2, predictors can be included in the model, where the intercepts or slopes of these predictors can vary across levels. Deciding whether to constrain or allow the level-1 and level -2 predictors to vary across clusters is an important aspect to MLM specification (Hong & Raudenbush, 2006; Lingle, 2009; Raudenbush & Byrk, 2002). For example, when data are nested, adjustment needs to be made for the cluster to ensure that individuals in very different clusters are not compared or matched to one another (Rosenbaum, 1986).

As individual units within a cluster tend to be similar, an intraclass coefficient (ICC) is used to assess the degree of dependency within a data set (O'Connell & McCoach, 2008). The ICC is calculated by decomposing the total variance into its within and between components and

results in a measure of the proportion of variance between the clusters (O'Connell & McCoach, 2008; Raudenbush & Bryk, 2002).

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \quad (14)$$

One way to resolve the challenges associated with multilevel data would be to consider the causal inference at the cluster level, where the clusters are treated as individual units. For example, Stuart (2007) investigated the effects of school-level data, by estimating the PSs using school level variables and aggregated individual student variables. However, the interpretation of findings from this method need to be clear as this method cannot be used to interpret relationships at a lower level. Often, some are inclined to erroneously apply relationships at a higher level to the lower level, also known as the “ecological fallacy” (Hox, 2002; Stuart, 2007). It would be inappropriate to assume if a new curriculum is shown to be effective at increasing the overall achievement of the school, that the curriculum will be effective for individual students.

Propensity Score Analysis with Multilevel Modeling

Multilevel research designs. The nature of the multi-level research design and structure of the assignment mechanism are important factors to consider in order to satisfy the assumptions of causal research. Hong and Raudenbush (2003) considered three different cross-sectional multi-level research designs: multi-site designs, cluster designs, and joint multi-level designs. Multi-site designs suggest that individuals within each cluster may be exposed to either treatment or control. In contrast, entire clusters are assigned to conditions in cluster designs. Lastly, in a joint multi-level design, clusters are assigned to a set of treatments, and within each cluster the level-1 units are assigned to a set of different treatments (Hong & Raudenbush, 2003).

Cluster level designs treat the organization or classroom or school as a single unit assigned to a treatment or control condition and the individuals within the cluster can be considered repeated measures of the cluster’s response to the treatment (Donner & Klar, 2000; Hong, 2004). Since the entire cluster is assigned to either treatment or control, potential bias is

associated with the observed cluster level covariates, W , and unobserved cluster level covariates, U_w (Hong & Raudenbush, 2003). Accordingly, treatment assignment at the cluster level, D , is independent of X , (level 1 covariates), U_x , (level 1 unobserved covariates) and the potential outcome, Y , given W and the U_w (Hong & Raudenbush, 2003). The conditional probability of assigning cluster j to the experimental condition is:

$$Q_j = \Pr(D_j = 1 | W_j, U_{wj}) \quad (15)$$

Therefore, Q is the probability that a cluster will be assigned to the treatment, which is constant for all individuals within said cluster (Hong, 2004; Hong & Raudenbush, 2003).

Estimation models. Given that the outcome is a probability, basic hierarchical models cannot be used; instead hierarchical generalized linear models (HGLM) need to be applied (O'Connell, Goldstein, Rogers, & Peng, 2008). As implied by its name, HGLMs are parallel to GLMs (e.g. logit models), within a hierarchical framework. Many of the implications regarding the nature and interpretation of single level logistic regression models apply to hierarchical logistic regression.

When estimating a binary outcome, such as a propensity score, in MLM, the level-1 residual is absent. The variance at the individual-level is determined by the mean of Z , which is equal to the probability of being in the treatment group. Thus, it is not a free parameter. However, within a hierarchical generalized linear framework several modeling choices can be made. That is, level-1 predictors can be assumed to be fixed across clusters or allowed to vary. Additionally, level-2 variables can be included to affect intercepts only or intercepts and slopes. Different models carry different levels of restrictions and implications regarding the assumptions about the selection process. An extremely restrictive MLM would be one in which the PSs are estimated with fixed slopes and no level 2 variables.

$$\begin{aligned}
\text{logit}(Z_{ij}) &= \beta_{0j} + \beta_{1j}X_{1ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10}
\end{aligned} \tag{16}$$

$$\begin{aligned}
\text{logit}(Z_{ij}) &= \gamma_{00} + \gamma_{10}X_{1ij} + \mu_{0j} \\
\mu_{0j} &\sim N(0, \tau_{00})
\end{aligned}$$

Here, the $\text{logit}(Z_{ij})$ represents the estimated PS, or the conditional probability of individual i in cluster j of receiving treatment. The γ_{00} represents the mean propensity score across classrooms while the γ_{10} is the mean contribution across classrooms of the individual-level predictor X_i to the PS. By fixing the slopes, it is assumed the effect of X_i on treatment assignment is constant across schools (Hong & Raudenbush, 2006). Additional models include allowing for cross-level interactions where the slopes and intercepts vary freely across clusters and allowing level 2 variables to effect intercepts and slopes. The decision on how to specify the level-1 and level-2 variables should be made based on the nature by which the cluster membership impacts the assignment mechanism is theorized (Hong & Raudenbush, 2003; Hong, 2004; Thoemmes & West, 2011).

Thoemmes and West (2011) introduced the idea of narrow and broad inference space, where narrow inference space aims to estimate PSs that mimic a randomized multisite trial while the broad inference space estimates PSs that mimic a randomized individual trial with incidental clustering and advise considering whether subject level or population wide inferences are desired when considering the estimation models. Specifically, if the narrow inference space is desired, researchers suggest including a full MLM model with fixed and random effects:

$$\begin{aligned}
\text{logit}(Z_{ij}) &= \beta_{0j} + \beta_{1j}X_{1ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + \mu_{1j}
\end{aligned} \tag{17}$$

combined as,

$$\text{logit}(Z_{ij}) = \gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + \mu_{1j})X_{1ij} + \mu_{0j}$$

Conversely, if the broad inference space is required then the model will be similar to that in Equation 16, omitting the random effects as these effects yield PSs based on different equations for each cluster (Thoemmes & West, 2011).

Conditioning methods. In MLM, PSs can be conditioned either within a cluster or across clusters. Often the nature of the research design and the estimation model dictates how the conditioning should occur. For example, in both multi-site and joint multi-level studies PSs may be conditioned within a cluster or between clusters, while cluster designs restrict conditioning of the estimated PSs between clusters.

Generally, conditioning within clusters is ideal (when feasible) as the variability between clusters is not integrated within the inferences (Lingle, 2009; Thoemmes, 2009; Thoemmes & West, 2011) and the original multilevel structure of the data is preserved which can be helpful when examining the variation of treatment effects across clusters (Kim & Seltzer, 2007). However, conditioning within clusters may potentially yield different results for the clusters, especially when cluster level variables are not included in the estimation model or are considered fixed (Thoemmes & West, 2011). Although the variability between clusters is introduced when conditioning across clusters, this method can be useful in certain designs when the clusters contain a small number of individuals (Arpino & Mealli, 2011). Thoemmes and West (2011) point out that when using a narrow inference estimation method, the estimated PSs will be based on different equations and therefore conditioning across clusters will most likely not yield balanced estimates. When conditioning across clusters, balance may not be achieved on individual covariates within each cluster, but rather the balance is on average achieved on the entire sample of covariates on all levels (Thoemmes & West, 2011). This difference can be likened to the differences between optimal matching and greedy matching, where optimal

matching considers the overall best set of matches for the entire sample, and greedy matching considers the best matches for each individual (Stuart, 2010).

Lastly, there are currently four methods to condition the PS within a single level context (matching, stratification, covariance adjustment, and weighting). However, in MLM, weighting is a bit more complicated (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). In a single level context, data are considered independent and therefore a simple weighting of the sample observations can be conducted. However, in MLM, the observations are no longer independent and therefore the method of weighting needs to be altered. Several researchers (e.g. Asparaouhov, 2004; Chantala, Blanchette, & Suchindran, 2011; Pfefferman et al., 1998) have considered different methods of weighting in MLM; however these methods are still novel and under development; therefore, integrating their application to PSs is premature.

Research on Propensity Scores in Multilevel Contexts

PS methods in hierarchical settings have received limited attention from both a methodological as well as an applied perspective. Rosenbaum (1986) recognized the challenges in applying PS methods to nested data and limited the PS estimation to individual variables and the matching of treatment and control individuals within clusters only. Specifically, he considered two estimation approaches for conditioning within clusters: including individual covariates and a binary covariate for each school, and ignoring school membership completely. The first approach was troublesome in Rosenbaum's (1986) study because as the number of clusters' increases the degrees of freedom would decrease and the dataset included 1,015 clusters. Hong (2004) pointed out similar issues with regard to the degrees of freedom may arise when trying to estimate the fixed effect of each school in cases where the number of individuals receiving treatment per cluster is small.

The second approach estimated the PSs without considering school membership. According to Rosenbaum, (1986) the between-cluster components of the variability were controlled by limited the matching to within clusters, and that both observed or unobserved fixed

cluster-level characteristics that affected treatment assignment would be balanced. Therefore, the treatment effects could be estimated without accounting for the clustering in the propensity score (Rosenbaum, 1986). However, this approach is limiting as in some cases matches may not be possible within clusters, or in cluster level designs where entire clusters receive the same condition.

More recently a select group of scholars have begun to apply PS methods to estimate the effects of different educational programs. For example, to estimate the effects of kindergarten retention on academic achievement and compare units across clusters, Hong and Raudenbush (2005, 2006), estimated the propensity of being retained using a hierarchical logistic model with random cluster effects (schools) and fixed slopes for 207 individual and 238 school level predictors. The fixed slopes assumed the effect of those individual predictors on treatment assignment was constant for individuals across schools (Hong & Raudenbush, 2006), which becomes an issue if cross-level interactions are present (Kim & Seltzer, 2007; Raudenbush & Bryk, 2002).

Kim and Seltzer (2007) compared PS estimation methods for a non-randomized multisite setting, where such a context may potentially lead to a large variability in the assignment mechanism between clusters. Using data from the Early Academic Outreach Program (EAOP) four multiple logistic regression estimation models were compared: a. single-level logistic regression model where only individual level predictors were included; b. multilevel logistic regression model with fixed slopes and random cluster level intercepts; c. multilevel logistic regression model with random intercepts and random slopes including both individual and cluster level predictors; and d. a multilevel logistic regression model with random intercepts and random slopes where cluster level predictors were not included (Kim & Seltzer, 2007). The second model explored Hong and Raudenbush's (2005, 2006) propensity score model using within cluster matching while the third and fourth models assumed cross level interactions of the

predictors. Allowing the intercepts and slopes to vary across clusters investigates whether the effects of the level-1 covariates differ across schools.

Individuals who participated in the EAOP were matched to individuals within their school not participating. Specifically, nearest neighbor matching without replacement with a caliper of ± 0.1 (standard deviation of the entire sample's PSs) was used. Overall, findings suggested that when conditioning within clusters, random effects with fixed slopes performed well. When considering random intercepts and random slopes, the omissions of level 2 predictors, or the random effects of the slopes had a significant impact upon the balance and the matched pairs. In contrast, including the random effects of the slopes, significantly improved the overall balance of the matched pairs (Kim & Seltzer, 2007). Both of the aforementioned studies used multilevel models to estimate PSs in applied settings. Since then, several researchers have begun to investigate the utility of PS estimation methods with MLM using Monte Carlo simulation methods (e.g. Arpino & Mealli, 2011; Lingle, 2009; Thoemmes, 2009; Thoemmes & West, 2011).

As their doctoral dissertations, both Lingle (2009) and Thoemmes (2009) examined the behavior of PS methods with hierarchical structures. Extending Kim and Seltzer's (2007) work these studies considered multi-site research contexts. Lingle's (2009) study evaluated three PS estimation models in their ability to achieve covariate balance across the sample as a whole and within each cluster, while Thoemmes (2009) evaluated the different PS estimation methods in their ability to estimate treatment effects.

To evaluate the balance achieved when using PSs with nested data, Lingle (2009) estimated random slopes and random effects MLM (see equation 18) and compared its results to the results from two single level models.

$$\begin{aligned}
\log it(Z_{ij}) &= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}W_1 + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}W_1 + \mu_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}W_1 + \mu_{2j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}W_1 + \mu_{3j}
\end{aligned} \tag{18}$$

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \end{pmatrix} \sim \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ \tau_{10} & \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{20} & \tau_{21} & \tau_{21} & \tau_{31} \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix}$$

The first single level model included all three level-1 predictors and the level-2 predictor which was collapsed to the individual level and the second single level model included all three predictors with no cluster level components (Lingle, 2009). The estimated PSs were conditioned using nearest neighbor matching within and between clusters and stratification. Specific data characteristics that were controlled included: individual and cluster level sample sizes, ratio of treatment to control group members, variable relationship to treatment assignment. The PS methods were fully crossed with the specific data characteristics. Subsequently, the methods simulated were applied to data available from the 2002 Educational Longitudinal Study and findings were compared to the simulation results to evaluate their applicability.

The performance of the propensity score models was determined across the sample as a whole and within each cluster. Covariate balance from each propensity score model and conditioning method per sample were compared. Balance was determined across the sample as a whole using the γ_{10} and the standardized mean differences of the pretreatment variables. Within each cluster balance was measured by determining the variance in β_{1j} using the values of τ_{11} . By removing the relationship between the treatment assignment and each covariate, a MLM that yields a non-significant relationship between the variables in treatment and control group implies the groups are balanced on the variable (Lingle, 2009).

Results did not indicate one method to outperform others given the set of conditions

simulated. Stratification across quintiles led to the largest number of treatment units retained followed by between group matching and then within group matching (Lingle, 2009). While within group matching retained the smallest number of treatment individuals, on average this method resulted in the fewest number of significant parameters. This could mean that the balance was greatest when matching within clusters, or that the loss of sample size and power made it difficult to elicit significant differences (Lingle, 2009). Additionally, between-cluster matching using propensity scores that are estimated using a logistic model, with or without the inclusion of a cluster-level predictor, resulted in very few significant differences across sample-size conditions. The variability across all conditions was the smallest with between group matching, and the greatest with stratification. Lastly, within cluster matching showed little differences among the three different estimation methods.

Similarly, Thoemmes (2009) and later Thoemmes and West (2011) explored the differences between the estimation and conditioning choices for estimating treatment effects in multisite randomized experiments under specific data characteristics. Thoemmes (2009) conducted a series of simulation studies in an attempt to identify the best MLM estimation model to use in multi-site studies. Compelling findings were later published in a special edition of *Multivariate Behavior Methods* (Thoemmes & West, 2011).

Specifically, four estimation models were examined: a. a single level model; b. fixed effects model; c. multilevel model with narrow inference space; and d. a multilevel model with broad inference space. Conditioning was conducted both within and across clusters. These PS techniques were fully crossed with two sample factors, (sample size and ICC of individual variables).each with two individual levels. Overall results indicated that single level models performed the worst, especially as the values of the ICC for the level 1 predictors increased (Thoemmes & West, 2011). With regard to bias, the multilevel model with narrow inference space conditioning within clusters performed marginally the best. Regardless of conditioning technique, the single levels models performed the worst and the remaining models were all

comparable with slightly larger bias in the models that conditioned across clusters. Models yielded similar results regarding the balance estimates. Overall these studies support conditioning within cluster when feasible.

To investigate the bias of the ATT estimators when using PS matching with unobserved cluster-level covariates Arpino and Mealli (2011) conducted a series of simulation studies investigating the specification of PSs with multilevel data structures common to large-scale survey research, where the size of the cluster is generally small. The purpose of this study was to investigate how a potentially relevant but unobserved cluster level variable can influence the overall results. Four different PS estimation models were considered, and results were compared when the cluster level variable was included as well as omitted. Specifically a benchmark single level logistic regression model with three level-1 predictors and the cluster level variable, a single level logistic regression model with three level-1 predictors, a MLM random intercepts model with three level-1 predictors, and a single level logistic regression model with dummy coded vectors to represent cluster membership. In addition each of these models was considered with a cluster mean of the level-1 variables as a substitute for the unobserved cluster variable (Arpino & Mealli, 2011). These simulations were repeated to compare different cluster sizes, different levels of potential outcome effects with and without the presence of cross level interactions and largely unbalanced data structures (Arpino & Mealli, 2011). Results indicated that when cluster level covariates are not collected, a fixed effects model may be the best method to exploit the hierarchical structure of the data as this model performed best across conditions (Arpino & Mealli, 2011).

Overall results showed omitting the cluster information did not perform as well as the model that included the cluster- level information with regard to the bias and MSE. In addition, the model with the cluster level dummy vectors achieved a high degree of balance. However, a high level of imbalance was found with the MLM model with the random intercepts. Overall,

researchers recommend using random and fixed effects models to capture any unobserved heterogeneity when a cluster level variable is unobserved (Arpino & Mealli, 2011).

In conclusion, the research on PS with MLM while continuing to thrive and develop is still fairly novel. The studies above focused on cross sectional clustered data, where individuals are nested within a second level. Additional studies, which are not of relevance to this dissertation, have examined PS methods with longitudinal data, where the hierarchical structure is based on a repeated measures design (Hughes, Chen, Thoemmes, & Kwok, 2010), complex samples (Hahs-Vaughn & Onwuegbuzie, 2006) and even incorporated measurement models into a three level analysis to assess the different sources of data (Hong & Yu, 2008).

The previously described preliminary studies introduced the challenges as well as the importance of considering the nested nature of the data structure when drawing casual inferences. While these studies have provided fruitful information, the body of knowledge regarding the performance of PSs within a multilevel framework remains sparse. Currently the focus of the work has been on multisite designs or designs where the clustering may be incidental. More empirical evidence for how to handle designs where the nested nature is deliberate needs some attention. Furthermore, the inclusion of larger numbers of individual and cluster covariates is needed to apply findings to educational data where large numbers of pretreatment variables are collected. Current research has based the performance of PSs methods on either balance estimates (Lingle, 2009) or the treatment effect estimates (Thoemmes & West, 2011). Examining these two outcomes mutually rather than exclusively may be worthwhile. Additionally, current investigations have examined a limited breadth of the PS conditioning methods. Finally, additional work comparing the differences among more of these conditions would be helpful to see if in different contexts certain techniques are more appropriate than others.

This dissertation aims to extend the empirical research on multisite designs by exploring various multilevel models to estimate PSs and multiple strategies to condition the estimated PSs across clusters, as well as increasing the complexity of the sample characteristics. The

performance of the different estimation models, conditioning strategies, and sample characteristics will be evaluated based on their ability to balance groups as well as estimate the treatment effect.

Chapter Summary

This literature review presented a myriad of theoretical and empirical literature regarding causal inference and PS. What is interesting to note is the empirical literature on PS in a single level is relatively far more advanced and includes many studies from the medical literature compared to the empirical literature on PS in hierarchical structures which is almost all based on educational or social and behavioral settings. With the continuous trend towards examining the effectiveness of educational programs additional generalizable and applicable evidence for this method is needed. This study aimed to investigate the performance of PS methods (both estimation models and conditioning methods) extended to a 2-level hierarchical context.

CHAPTER THREE: METHOD

This chapter outlines the proposed methodology for this study and includes a description of the purpose, research questions, design, sample characteristics, data generation methods, analytical procedures, and outcome measures.

Study Purpose

The purpose of this study was to further examine the appropriateness of using PS methods to achieve balance between groups on observed covariates and to yield unbiased treatment effect estimates in multilevel studies. Specifically, this study examined the extent to which different PS methods (PS estimation models and PS conditioning strategies) and sample characteristics (sample size, strength of covariate associations to assignment and outcome, parameters predicting treatment assignment, and population effect size) achieve balance and reproduce the population treatment effect. PSs were specified using four different logistic regression models and subsequently conditioned using three different strategies across clusters.

PS estimation models included both a single level model and three multilevel models: (a) single level (b) random intercepts, (c) random coefficients and, (d) cross level. For each of the four PS estimation models, three different PS conditioning strategies were investigated and included: (a) matching, (b) stratification, and (c) covariance adjustment. PS methods (estimation models and conditioning strategies) fully crossed with several sample characteristics were compared to evaluate the quality of balance achieved and the accuracy and precision of treatment effect estimates produced.

Research Questions

1. To what extent do balance estimates vary across PS methods (PS estimation models and PS conditioning strategies)?

2. To what extent do data factors (level 1 and level 2 sample sizes, strength of covariate associations to assignment and outcome, and population effect size) affect the balance achieved by the PS methods (PS estimation models and PS conditioning strategies)?
3. To what extent do treatment effect estimates vary across PS methods (PS estimation models and PS conditioning strategies)?
4. To what extent do data factors (level 1 and level 2 sample sizes, strength of covariate associations to assignment and outcome, and population effect size) affect the treatment effects estimated by the PS methods (PS estimation models and PS conditioning strategies)?
5. What is the direction and strength of the relationship between balance and both the accuracy and precision of treatment effect estimates?

Design

This study incorporated a 4 x 3 x 3 x 3 x 2 x 2 x 2 x 2 x 2 factorial design. The following nine independent variables were included: (1) PS estimation models (single level, random intercepts, random coefficients, and cross-level); (2) PS conditioning strategies (matching, stratification, and covariance adjustment); (3) number of clusters (small [n=30], moderate [n=50], and large [n=100]); (4) within-cluster sample size (small [n =01-09], moderate [n =10-19], and large [n =20-29]); (5) relationship between level-1 covariates and treatment assignment (small [$\beta_{xz}=.1$], and moderate [$\beta_{xz}=.2$]); (6) relationship between level-1 covariates and outcome (small [$\beta_{xy}=.1$], and moderate [$\beta_{xy}=.2$]); (7) relationship between level-2 covariates and treatment assignment (small [$\gamma_{wz}=.2$], and moderate [$\gamma_{wz} = 0.4$]); (8) relationship between level-2 covariates and outcome (small [$\gamma_{wy}=0.2$], and moderate [$\gamma_{wy} =0.4$]); and (9) population effect size (δ = small [0.2] and moderate [0.5]). All levels of all the factors were fully crossed with one another for a total of 3,456 conditions. Data were generated using SAS software (version 9.2 and 9.3; SAS Institute, 2008) through the IML procedure. For 256 of the 288 design cells, 1000 datasets were generated, for the 32 conditions where the number of level 1 units was 20-29 and the number of clusters was 100, 500 datasets were generated. Given the complexity of the various analytical

procedures, the conditions with the large number of units within a large number of clusters proved to be both very computer and time intensive; therefore fewer datasets were generated for these conditions. A total of 272,000 datasets were generated using the IML procedure in SAS and subsequently analyzed using different PS estimation models and conditioning methods.

Samples

The samples for this study were based on a two-level hierarchical model in which individuals units, i , are nested in clusters, j . At the first level, a continuous outcome (Y) was generated as a linear function of 30 (27 continuous and 3 dichotomous) predictors, X and one binary assignment variable, Z (see equation 19). The cluster level was simulated with 10 (9 continuous and 1 dichotomous) level-2 predictors W . The intercepts and slopes of five X (4 continuous and 1 dichotomous) level 1 predictors varied randomly across clusters. In addition 1 continuous cluster level variable interacted with three level 1 continuous fixed variables. For brevity, common statistical notations are used to present the equations at each level. The expanded version, specifying each parameter individually can be found in Appendix A.

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \dots + \beta_{30j}X_{30ij} + \beta_{Zj}Z_{ij} + e_{ij}$$

where,

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1j} + \dots + \gamma_{010}W_{10j} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \mu_{1j} \\ &\vdots \\ \beta_{5j} &= \gamma_{50} + \mu_{5j} \\ \beta_{6j} &= \gamma_{60} \\ \beta_{7j} &= \gamma_{70} \\ \beta_{8j} &= \gamma_{80} + \gamma_{81}(W_{1j}) \\ \beta_{9j} &= \gamma_{90} + \gamma_{91}(W_{1j}) \\ \beta_{10j} &= \gamma_{100} + \gamma_{101}(W_{1j}) \\ \beta_{11j} &= \gamma_{110} \\ &\vdots \\ \beta_{30j} &= \gamma_{300} \\ \beta_{Zj} &= \gamma_{Z0} \end{aligned} \tag{19}$$

combined as;

$$\begin{aligned}
Y_{ij} &= \gamma_{00} + \gamma_{Z0}Z_{ij} + \sum_{s=1}^{10} \gamma_{0s}W_{sj} + \sum_{s=1}^{30} \gamma_{s0}X_{sij} + \gamma_{81}(X_{8ij})(W_{1j}) + \gamma_{91}(X_{9ij})(W_{1j}) + \gamma_{101}(X_{10ij})(W_{1j}) + \\
&\mu_{0j} + \mu_{1j}X_{1ij} + \mu_{2j}X_{2ij} + \mu_{3j}X_{3ij} + \mu_{4j}X_{4ij} + \mu_{5j}X_{5ij} + e_{ij} \\
e_{ij} &\sim N(0, \sigma^2) \\
\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ & \tau_{11} & & & & \\ & & \tau_{22} & & & \\ & & & \tau_{33} & & \\ & & & & \tau_{44} & \\ & & & & & \tau_{55} \end{pmatrix} \right)
\end{aligned}$$

The large numbers of individual and cluster level predictors were intended to mirror common educational data structures. Previous simulation studies have incorporated only a few covariates (e.g. Arpino & Mealli, 2011; Lingle, 2009, Thoemmes, 2009; Thoemmes & West, 2011). However, in their review of the studies applying PS methods, Thoemmes & Kim (2011) found that researchers often use many covariates to estimate PSs. Specifically, for the 88 studies analyzed, a mean of approximately 30 covariates was reported (Thoemmes & Kim, 2011). The major benefit for utilizing PS methodology over traditional matching methods/statistical adjustments is the ability to reduce a large number of pretreatment covariates to a single scalar. Rarely are few covariates adequate to satisfy the assumption of strongly ignorable treatment assignment. When only a few covariates are used, it may be prudent to match using the actual covariates as they function as a much finer balancing score than the PS which is the coarsest balancing score (Rosenbaum & Rubin, 1983). Thus, the inclusion of a large number of predictors intended to simulate data commonly found in applied research.

Allowing the level-2 variables to influence the slopes and intercepts of five level-1 predictors builds upon prior research. Lingle (2009) modeled one level-2 predictor to influence the intercepts and slopes of all three of the level-1 variables. All variance and covariance parameters in the matrix of random effects were estimated. In contrast, Thoemmes (2009) most

complex model included nine level-1 continuous variables, of which the slopes and intercepts for two predictors varied across level-2 units. No level-2 predictors were included. Additionally, modeling dichotomous variables at either level has not previously been considered.

In order to adequately evaluate the simulated conditions, several simplifying assumptions were imposed on the samples. First, all continuous predictors were standardized with a mean of 0 and a variance of 1.0. Dichotomous predictors were generated from a binomial distribution where the population mean is approximately 0.5, which translates to an odds ratio of 1, or an equal probability for each group. Next, assignment to treatment, Z , is a binary grouping variable generated at the individual level using the RANUNI random number generator in SAS version 9.2, where 25% of the individuals in the cluster were assigned to treatment and the remaining 75% to the control group. This imposed a 1:3 treatment to control ratio. This ratio was used by Gu and Rosenbaum (1993) and Lingle (2009) and represents moderate difficulty when matching on the PS is used (Gu & Rosenbaum, 1993).

The fixed effects for the intercepts were set to 1.0. The level-1 errors were generated from a normal distribution with a variance of 1.0 using the RANNOR random number generator in SAS version 9.2 (SAS Institute, 2008). The level-2 errors were also generated from a normal distribution with a variance of .25 to produce conditional ICCs of .20 for the predictors with varying slopes. An ICC of .20 is substantially large and implies a great deal of dependency in the data; however, is not out of the range of typical ICCs found in school effects research (McCoach, 2010). The correlations among the predictors at each level were held constant at 0.2. The variables in both Lingle (2009) and Thoemmes (2009) study were assumed to be uncorrelated, which is fairly unrealistic in social science research. Lastly, similar to Thoemmes (2009) samples, data were generated using a variance components covariance structure where all covariance parameters in the random effects matrix were fixed to 0. This was done to reduce the number of parameters to estimate in a true-confounders model. The current study incorporated several complex data factors (i.e. the increased number of covariates, the use of correlated

variables, the inclusion of level 2 predictors, the addition of dichotomous variables and the different sample size conditions which produce smaller samples) that had not been addressed in the literature, yet are common characteristics found in social science research. In light of the infancy of empirical investigations of this nature it seems prudent to first begin to understand the behavior of the methods under only a few complex factors maintaining some simplifying conditions.

The data generation approach for producing correlation matrices and values for variables was empirically driven for each individual sample based on the eigenvalues of the population correlation matrices. When positive eigenvalues are found, the sample was generated using the Cholsky decomposition approach; however, if one or more non-positive eigenvalues are present, the sample was generated using the Principle Component Analysis (PCA) approach (Fan, Felsovalyi, Sivo, & Keenan, 2002).

These aforementioned assumptions and procedures remained constant across the different combinations of sample characteristics simulated. These sample characteristics function as the independent variables for this study and are described in detail below.

Sample Characteristics

Sample size. Three different values for the total number of clusters were examined (30, 50, and 100). These cluster sizes were chosen to align with Lingle's (2009) clusters and are consistent with applications of MLM in educational research (Dedrick et al., 2009). The sample size within each cluster varied and was assigned from a uniform distribution. This design results in datasets with clusters varying in size, which is what would realistically be found in substantive research.

The number of individuals nested within each cluster varies to represent small, moderate, or large clusters. Small clusters ranged from 01-09, moderate from 10-19, and large from 20-29. These samples sizes were randomly generated and uniformly distributed. Clusters of only 01-09 individuals are smaller than the general guidelines set regarding level-1 sample size; however

clusters with few members is not uncommon in social science research (Bell, Ferron, & Kromrey, 2008). For example, clusters with less than 10 individuals would not be uncommon in cases where treatment may be rare or conducted on special populations. Often the estimation of causal effects for different subpopulations is of interest in educational research (NCLB, 2002). Therefore if treatment is administered to special populations at each cluster, a smaller sample size within cluster would be anticipated. In contrast, the larger cluster size was chosen to accommodate studies of whole classrooms as well as other programs such after-school tutoring programs where assignment to treatment may be self-selected. In addition, these sample sizes are comparable to those examined by Lingle (2009) Thoemmes (2009) (i.e. 10, 30, 25, 50 and 100 clusters). One extension from the previous research was the use of a range of sample sizes to represent small, moderate, and large clusters of individuals to increase the generalizability to educational settings where the sizes of clusters will likely vary.

Relationship between covariates and treatment assignment. To satisfy the strongly ignorable treatment assignment assumption, predictors associated with assignment mechanism should be controlled. Previous research suggests that predictors that are marginally associated or unassociated with assignment can tend to increase the variance and reduce the efficiency of the estimates (Austin, Grootendorst, & Anderson, 2007; Austin, Grootendorst, Normand, & Anderson 2007; Brookhart, et al., 2006; Pearl, 2010; Wooldridge, 2009). Therefore, values representing small and moderate relationships were considered. Given the unrealistic nature of having a completely uncorrelated predictor, a magnitude of 0 was not chosen. Thoemmes (2009) simulated the data sets based on values representing the explained variability of all the covariates. Given that the predictors were uncorrelated the resulting path coefficients could be interpreted as correlations. This study resulted in slopes of .047 and .169 representing small and large effects (Thoemmes, 2009). In contrast, Austin and associates' studies (2007a, 2007b) as well as Lingle's (2009) simulation used varying magnitudes representing the strength of the covariate relationship

to treatment including 0. None of these studies included either correlated or dichotomous predictors.

In order to select realistic values to represent small and moderate relationships between predictors and treatment assignment preliminary exploratory simulations were conducted using brute force techniques to obtain population parameters. A sample of 1000 clusters containing between 500 and 1000 units was simulated based on the assumptions described above. True propensity scores for the individual units were calculated using different values as the regression weights. Covariates were generated from a normal distribution, where the odds of receiving treatment is a function of each covariate, and the conditional likelihood for all the covariates combined, producing a probability of .25. This was done by setting the log odds, or the intercept value for the probability equation to -1.108. A correlation of .2 was induced among the variables at each level. The individual units in the clusters were assigned to either treatment or control based on their true propensity scores. Next, both correlation matrices and tolerance values were examined to ensure the magnitude of the slopes were realistic.

Values of .1 and .2 yielded small correlations between treatment assignment and the level 1 and level 2 predictors respectively, while values of .2 and .4 produced moderate correlations. The correlations among the predictors at each level remained approximately .2 with some slight variations for the dichotomous variables. The relationship between variables across levels stayed very close to 0 and any non-zero estimate did not seem to pose a concern.

One significant contribution of this study is the use of correlated predictors. Previous studies have examined the behavior of PS estimates with perfectly uncorrelated predictors—an unlikely and rare phenomena. However, correlated independent variables have the potential to adversely impact regression estimates to the extent that they are no longer interpretable (Pedhauzer, 1997). Pedhauzer (1997) illustrates how the variance of a regression coefficient can become inflated when correlated independent variables exist. In addition, when multiple independent variables are present, simply assessing the collinearity based on the zero-order

correlations is insufficient as the variance inflation factor may still be high with relatively lower zero-order correlations.

To assess the degree of collinearity among the variables, R^2 values for fitted ordinary least squares (OLS) regressions based on the four combinations of population correlation matrices were computed for each predictor. Each predictor was modeled against the remaining predictors to obtain its individual R^2 value. For the purpose of obtaining R^2 values to examine each predictor's unique contribution, dichotomous variables (including Z) were treated as continuous and were fitted using OLS rather than a logit or probit models as they do not produce comparable R^2 values. Tolerance, is defined as $1 - R_p^2$, for each predictor against remaining predictors. Small tolerance values suggest greater adverse effects due to the collinearity among the variables. While there is no agreement as to what constitutes acceptable tolerance, general guidelines do exist (Pedhauzer, 1997). Some statistical programs use a default value of .01, and variables with tolerance levels less than .01 are excluded from the analysis. R^2 values ranged from .0940 to .488. None of these values suggest questionable tolerance values; therefore the magnitude of the regression weights predicting treatment assignment were simulated using values of .1 and .2 for the level 1 predictors and .2 and .4 for the level 2 predictors.

Although this study intended to deliberately impose a relationship among the predictors, higher tolerance values suggest each variable contributes uniquely to the variance in Y . While the goal of estimating PSs is not to maximize the proportion of explained variance parsimoniously (Shadish & Steiner, 2010), including variables that are not highly tolerable may increase noise with regards to the treatment effects (Brookhart et al., 2006; Wooldridge, 2009).

Relationship between covariates and outcome. Previous research on PS in a single level found that including variables related to outcome influenced the estimates (e.g. Austin, Grootendorst, & Anderson, 2007; Brookhart, et al., 2006; Kelcey, 2011; Wooldridge, 2009). Therefore, varying magnitudes representing small and moderate relationships between each

variable and the overall outcome were considered. In order to maintain consistency within the study, the same regression weights used to represent the relationship between covariate and treatment were also used to represent the relationship between covariate and the outcome. Thus, values of .1 and .2 represented small and moderate values for the level 1 regression coefficients, γ_{s0} , while values of .2 and .4 represented small and moderate values for the level 2 regression coefficients, γ_{0s} . Analogous to the estimation models, these values were also investigated using preliminary simulation research. Previously, R^2 values were assessed for each of the individual predictors. To test the regression weights for the outcome model, covariance algebra was applied based on the four population correlation matrices to produce population OLS R^2 values, or, the overall proportion of variability explained by all the predictors and treatment assignment. These values, which can be found in Appendix B, were used to evaluate whether the values for γ_{s0} , and γ_{0s} , were pragmatic with respect to educational research and did not consider the nested nature of the data. Therefore, they should be interpreted as approximations.

Population treatment effect. Two population treatment effect values were considered, representing situations where the treatment has small ($\delta=0.2$) and moderate ($\delta=0.5$) effects on the outcome (Cohen, 1988).

Analytical Procedures

This dissertation aimed to explore the performance of both PS estimation models and PS conditioning techniques within a MLM framework. Crossing four PS estimation models with three PS conditioning techniques, a total of 12 PS approaches was examined under each of the 288 combinations of sample characteristics thereby yielding 3,456 simulated conditions. Of particular interest, was the degree to which different PS methods were able to create balanced groups and reproduce the population treatment effect. To answer the research questions and draw inferences regarding the application of PS in MLM several sequential analytical procedures were conducted on the generated samples. Samples were created in IML using SAS version 9.2 or 9.3 and sent to Base SAS version 9.2 or 9.3 (SAS Institute, 2008) for all subsequent analysis.

Throughout the analysis, various descriptive measures and outcomes related to balance and treatment effects were collected and aggregated across samples. Details of these procedures and outcomes are illustrated in Figure 2.

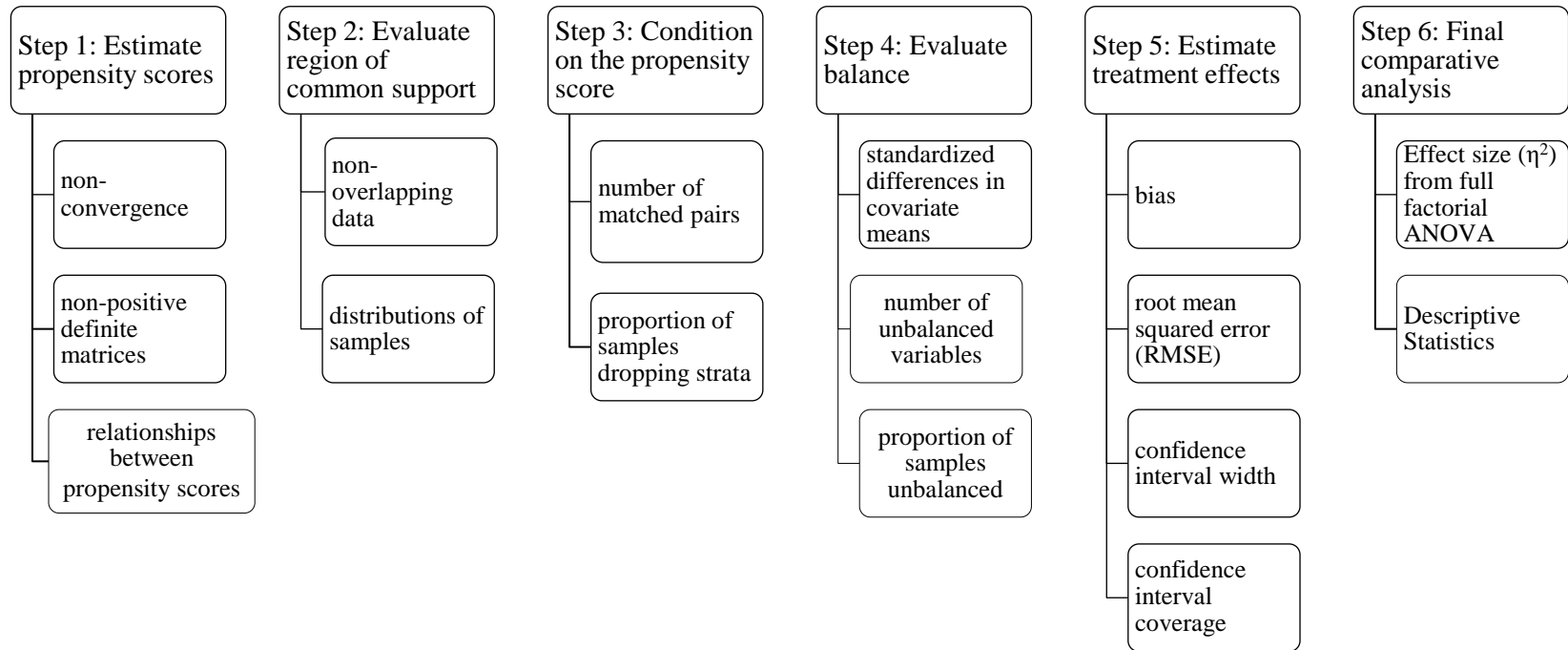


Figure 2. Analytical procedures with corresponding outcome measures

Step one: Propensity score estimation. For each of the 288 combinations of sample conditions, four different logistic regression models estimated the propensity scores on each replication using either the LOGISTIC or GLIMMIX procedure in SAS version 9.2 or 9.3 (SAS Institute Inc., 2008). The GLIMMIX procedure offers many useful options for fitting GLMs, including HGLMs (Dai, Li, & Rocke, 2006; Schabenberger, 2005). The four models chosen to estimate the PSs in this study paralleled the models Kim and Seltzer (2007) tested. Each model built upon the previous model and is presented in order of complexity using statistical notations to summarize the parameters. Expanded versions, including all parameters for each of the PS estimation models are appended (see Appendix C).

The first model, represented by equation 20, is a single-level logistic regression and represents situations where the nested of the data are ignored when estimating individual propensity scores. This model provided information regarding the utility of PS methodology for research applications which do not consider the effects of clustering. In order to adhere to best practices of PS methodology all level-1 and level 2 covariates were included. This model is referred to as the single level logistic or SLL model and is defined as follows:

$$\widehat{\text{LogitPS}}_{ij} = \beta_{00} + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{30} X_{30} + \beta_{31} W_1 + \dots + \beta_{40} W_{10} \quad (20)$$

The second model, equation 21, estimates the propensity score using a basic hierarchical random intercepts model. Here the multilevel nature of the data is accounted for by adding design variables to the single level model so that each cluster has its own unique intercept in the model. The cluster intercepts (subject-specific intercepts) are measured by the random intercepts, β_{0j} , which is a linear combination of the grand intercept, γ_{00} , the effects of the level-2 predictors ($W_{1j} - W_{10j}$) and a deviation, μ_{0j} from that intercept. The random intercepts are used to measure the differences between clusters, controlling for the effects of the individual predictors, X, which are considered fixed across clusters. Such a model, explored the performance of PSs in situations

where the clustering may be considered incidental. This model is referred to as the random intercepts model, or RI model.

$$\widehat{\text{logitPS}}_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \dots + \beta_{30j}X_{30ij}$$

where,

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \gamma_{03}W_{3j} + \gamma_{04}W_{4j} + \gamma_{05}W_{5j} + \gamma_{06}W_{6j} + \gamma_{07}W_{7j} + \\ &\gamma_{08}W_{8j} + \gamma_{09}W_{9j} + \gamma_{010}W_{10j} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} \\ &\vdots \\ \beta_{30j} &= \gamma_{300} \end{aligned} \tag{21}$$

combined as,

$$\begin{aligned} \text{logitPS}_{ij} &= \gamma_{00} + \gamma_{01}W_{1j} + \dots + \gamma_{010}W_{10j} + \gamma_{10}X_{1ij} + \dots + \gamma_{300}X_{30ij} + \mu_{0j} \\ \mu_{0j} &\sim N(0, \tau_{00}) \end{aligned}$$

The third model, referred to as the random coefficient model, allowed the slopes of five predictors to vary across clusters along with the random intercepts. Here, the effect of clustering is accounted for by allowing each level-2 unit to have its own unique intercept as well as its own unique slope for five of the level 1 predictors. As in the previous model, in this model the cluster intercept is a linear combination of the grand intercept, γ_{00} , the effects of level-2 predictors (W_{1j} - W_{10j}), and a deviation, μ_{0j} from that mean. Similarly, the random slopes, such as β_{1j} is specified as a linear combination of the overall average slope, γ_{10} and the cluster specific deviation from the overall average slope, μ_{1j} (see equation 22). This model aimed to replicate situations that consider all the substantive effects of the clustering at the individual level. This model is vital in order to address whether subject-specific effects (i.e. random intercepts and random slopes) adequately absorb the impact of the omitted cross-level predictors. This model represents scenarios where the model is correctly specified in terms of the covariate selection, but

the functional form of the model is incorrect (i.e. no cross-level interactions). This model is referred to as the random coefficients model or the RC model.

$$\widehat{\text{logitPS}}_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \dots + \beta_{30j}X_{30ij}$$

where,

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \gamma_{03}W_{3j} + \gamma_{04}W_{4j} + \gamma_{05}W_{5j} + \gamma_{06}W_{6j} + \gamma_{07}W_{7j} + \\ &\gamma_{08}W_{8j} + \gamma_{09}W_{9j} + \gamma_{010}W_{10j} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \mu_{1j} \\ \beta_{2j} &= \gamma_{20} + \mu_{2j} \\ \beta_{3j} &= \gamma_{30} + \mu_{3j} \\ \beta_{4j} &= \gamma_{40} + \mu_{4j} \\ \beta_{5j} &= \gamma_{50} + \mu_{5j} \\ \beta_{6j} &= \gamma_{60} \\ &\vdots \\ \beta_{30j} &= \gamma_{300} \end{aligned} \tag{22}$$

combined as,

$$\begin{aligned} \text{logitPS}_{ij} &= \gamma_{00} + \gamma_{01}W_{1j} + \dots + \gamma_{010}W_{10j} + \gamma_{10}X_{1ij} + \dots + \gamma_{300}X_{30ij} + \mu_{0j} \\ &+ \mu_{1j}X_{1ij} + \mu_{2j}X_{2ij} + \mu_{3j}X_{3ij} + \mu_{4j}X_{4ij} + \mu_{5j}X_{5ij} \\ \begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ & \tau_{11} & & & & \\ & & \tau_{22} & & & \\ & & & \tau_{33} & & \\ & & & & \tau_{44} & \\ & & & & & \tau_{55} \end{pmatrix} \right) \end{aligned}$$

Finally the last model, referred to as the cross-level model, includes all the true confounders at both levels, as specified by the data generation process. This model, (equation 23) builds upon model 3 the random coefficients model, by including the three cross-level interactions. This cross level model is identical to the random coefficients model, which included

random intercepts and five random slopes but included three additional fixed cross-level interaction effects. This model is referred to as the cross-level model or the CL model.

$$\widehat{\text{LogitPS}}_{ij} = \beta_0 + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \dots + \beta_{30j}X_{30ij}$$

where,

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}W_{1j} + \dots + \gamma_{010}W_{10j} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \mu_{1j} \\ &\vdots \\ \beta_{5j} &= \gamma_{50} + \mu_{5j} \\ \beta_{6j} &= \gamma_{60} \\ \beta_{7j} &= \gamma_{70} \\ \beta_{8j} &= \gamma_{80} + \gamma_{81}(W_{1j}) \\ \beta_{9j} &= \gamma_{90} + \gamma_{91}(W_{1j}) \\ \beta_{10j} &= \gamma_{100} + \gamma_{101}(W_{1j}) \\ \beta_{11j} &= \gamma_{110} \\ &\vdots \\ \beta_{30j} &= \gamma_{300} \end{aligned} \tag{23}$$

combined as,

$$\begin{aligned} \text{LogitPS}_{ij} &= \gamma_{00} + \sum_{s=1}^{10} \gamma_{0s}W_{sj} + \sum_{s=1}^{30} \gamma_{s0}X_{sij} + \gamma_{81}(X_{8ij})(W_{1j}) + \gamma_{91}(X_{9ij})(W_{1j}) + \gamma_{101}(X_{10ij})(W_{1j}) \\ &\quad + \mu_{0j} + \mu_{1j}X_{1ij} + \mu_{2j}X_{2ij} + \mu_{3j}X_{3ij} + \mu_{4j}X_{4ij} + \mu_{5j}X_{5ij} \\ \begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ & \tau_{11} & & & & \\ & & \tau_{22} & & & \\ & & & \tau_{33} & & \\ & & & & \tau_{44} & \\ & & & & & \tau_{55} \end{pmatrix} \right) \end{aligned}$$

The cross-level interaction model explored the performance of propensity score methods in situations where all the potential confounders are included in the model and is crucial in

assessing the utility of extending the current recommended methods of propensity score methodology to multi-level models. This model was the closest to the data generation model.

When using regression models to predict outcomes, one goal is to estimate models parsimoniously using the best set of predictors (Recchia, 2010; Shadish & Steiner, 2010). This is true for all predictive analyses, including least squares estimation or generalized linear model in both single and multi-level contexts (Recchia, 2010). However, when extending these models to estimate PSs, parsimony is no longer a goal, but rather viewed as a potential limitation. As previously mentioned, the inclusion of many and all predictors is necessary to estimate a single score representative of all confounders (Rosenbaum & Rubin, 1983; Shadish & Steiner, 2010; Steiner, Cook, & Shadish, 2011; Stuart, 2010). However, estimation issues often arise in multi-level models with random effects (i.e. random intercepts or random coefficients).

One concern for the PS estimation models, specifically the three multi-level models was whether these models would successfully converge and produce PS estimates. Both multi-level models as well as GLM's estimate effects using ML estimation (Agresti, 1996; Hox, 2002; O'Connell & Rivet Amico, 2010; Quinn & Keough, 2010). ML estimation incorporates an iterative procedure to estimate effects, and there is no guarantee that the iteration procedure will stop (Hox, 2002). If no solution is found before the computer program's set limit, then the model has not converged. The maximum number of iterations can be increased, but generally, after a large number of iterations if convergence has not occurred, it is possible that it may never find a solution. In multi-level modeling non-convergence is usually a sign of a poorly specified model, a very small sample size, or when many random variance components that are actually close to zero are being estimated (Hox, 2002). Although none of the estimation models try to estimate random variance components for fixed effects, convergence was still a valid concern; therefore non-convergence rates were tracked across models and sample conditions.

Thoemmes (2009) included a random intercepts model, and two random coefficients models, one with one random slope and the second with two random slopes. None of those

models encountered any convergence issues; however, in addition to the increased number of random slopes, the data in the current study have different characteristics that may lead to some sample non-convergence. Some major differences include the use of correlated variables, the inclusion of level 2 predictors, and the different sample size conditions which produce smaller samples. Therefore, the convergence status option was requested for each estimation model in order to gather descriptive information. The proportion of non-convergence for each condition was calculated and included as a descriptive statistic. Additionally, all subsequent calculations and comparisons were adjusted to represent the results based on the number of converged samples and not the number of samples generated. Finally, all interpretations and conclusions are limited to the samples successfully converging.

In addition, since all PS estimation models specify a variance components covariance structure, where all the covariance elements in the τ matrix are set to zero and only the variance components are estimated (Kincaid, 2005), it is likely to experience samples with non-positive definite matrices (Thoemmes, 2009). Even though eigenvalues were evaluated during data generation and negative variance estimates will be set to small positive values, the variance estimates may fall below zero when estimating PSs. In PROC GLIMMIX, when SAS encounters a negative variance, it replaces the value for a zero and issues a warning. Replacing the negative variance with a zero is common and seldom problematic if the estimated negative variance is small (Thoemmes, 2009). One way to potentially avoid issues with non-positive definite matrices would be to specify a different covariance structure (Kincaid, 2005); however, given the lack of empirical investigations for using PS methods in MLM, and the potential for issues of non-convergence, this option was not pragmatic for the current study.

Thoemmes (2009) encountered a substantial amount of non-positive definite matrices in models, especially conditions with small samples and random slopes. To ensure the results were not impacted by the presence of non-positive definite matrices, additional tests were conducted. For one condition a subsample of replications were re-estimated to examine the actual negative

variance values by removing the zero lower bound. The values for the negative variances for the sub-sample were quite small and were found to be within the 95% confidence interval; therefore it did not seem prohibitive to allow SAS to set the non-positive variances to 0. The estimated treatment effects between samples with and without non-positive definite matrices were then compared graphically as well as statistically. Overall results indicated estimates on average were not systematically different; in some instances significant differences were present, but small in magnitude. Additionally, these differences were not consistent across conditions. Findings were based on highly unbalanced samples, therefore it was tentatively concluded that setting a non-positive variance to zero would not influence the overall results (Thoemmes, 2009).

Given these provisionally favorable results all samples were generated with positive definite matrices, and any non-positive variance values estimated was automatically set to zero. Additionally, the proportion of samples producing non-positive definite matrices was computed across the different sample conditions for each estimation model.

Step two: Evaluating the region of common support. The region of common support, defined as the region of overlap between the estimated PS distributions for treatment and control units, provides a good measure for assessing and describing model accuracy (Shadish & Steiner, 2010; Stuart, 2010; Thoemmes & Kim, 2011). It is implied that causal effect estimates are to be considered for the units whose estimated PSs fall in the common support region (Shadish & Steiner, 2010; Stuart, 2010); thus a small region of common support is considered restrictive while a broad region of common support is preferred (Thoemmes & Kim, 2011).

Small common support regions could be due to a number of factors including model misspecification or violations to the strongly ignorable treatment assignment assumption.

In accordance with the implications regarding casual effect estimates and the region of common support, it is suggested that individuals with PSs outside of the region of common support be dropped from the analysis (Shadish & Steiner, 2010; Steiner & Cook, 2013; Stuart, 2010); however, this step is often overlooked or has been unreported in both the applied and

methodological investigations of PSs. For example, Lingle's (2009) study focused solely on the ability of PS methods to remove the selection bias by creating balance in treatment and control groups, yet there was no evidence to suggest the distributions of the PS were assessed let alone used to retain finer samples. Additionally, Thoemmes (2009) examined the distributions, but did not report any measures describing the region of common support, nor were out of bounds units dropped from the analysis.

In order to add to the current research, and adhere to the suggested guidelines provided in the literature, data were trimmed so that cases with estimated PSs that fall outside of the region of common support were dropped. The resulting data sets are thought to yield more comparable treatment and control groups at the expense of data reduction. To assess the degree to which trimming helped to create finer groups, the proportion of non-overlapping data were examined and the distributions in the samples before and after trimming were computed.

Step three: Propensity score conditioning. Current investigations of PS in MLM has incorporated either matching (Arpino & Mealli, 2011) or stratification techniques (Thoemmes, 2009; Thoemmes & West, 2011), with little emphasis towards comparing different conditioning methods (Lingle, 2009). Asymptotically, all conditioning methods should yield similar results (Rosenbaum and Rubin, 1983; Abadie and Imbens, 2006); however, currently there is not enough evidence to render judgments regarding which conditioning methods will yield the most favorable results for different data conditions when PS methods are applied in a single level (Shadish & Steiner, 2010; Steiner & Cook, 2013). Preliminary analyses suggests sample size may influence which conditioning method to use, and currently the literature describes the choice in conditioning method as a trade-off between bias and standard errors (Brookhart, et al., 2006; Caliendo & Kopenining, 2008; Pearl, 2010; Wooldridge, 2009).

In addition to technique, conditioning PSs in MLM modeling is also dependent upon whether or not the conditioning was conducted within clusters or across clusters. Current research favors conditioning within cluster to avoid introducing the variability across clusters, as

well as to ensure the units are all estimated using the same estimation equation (Lingle, 2009; Thoemmes, 2009; Thoemmes & West, 2011); yet all acknowledge that under certain models and sample conditions conditioning across clusters cannot be avoided and in certain contexts will not yield appreciably different results. Often in educational research, conditioning across clusters is preferred. For example, if the effects of an intervention conducted on special populations are desired, conditioning within cluster may not be feasible. Additionally, conditioning within clusters restricts the generalizability of the intervention to the cluster level while conditioning across clusters increases the external validity. Variations of three different conditioning techniques were conducted on each sample's estimated PSs. Recall each sample estimated four different PSs, one for each of the estimation models; thus each of the estimated PSs was conditioned three times. Specifically, each sample was conditioned *across* clusters using (a) nearest neighbor matching algorithm without replacement which produced 1:1 matched pairs with a caliper, or difference in estimated PSs no greater than .25, (b) ranks to stratify the distribution of estimated PSs into five sub-samples or strata, and (c) the estimated PS as a covariate in the multi-level outcome model.

This matching method was chosen for its feasibility with simulation research as opposed to the other more computer intensive algorithms and its common use in applied and simulation research (Gu & Rosenbaum, 1993; Ming & Rosenbaum, 2001; Thoemmes & Kim, 2011). Stratifying samples on quintiles was chosen as this is the most commonly advised number of subsamples to create (Rosenbaum & Rubin, 1984). While in practice, some studies have stratified finer subsamples (e.g. Hong & Yu, 2008), quintiles have been shown to at a minimum remove 90% of the bias in the variables used to estimate the PS (Rosenbaum & Rubin, 1983). Since, both Lingle (2009) and Thoemmes (2009) stratified samples into quintiles; findings can be compared across studies to determine the degree of consistency for this conditioning method.

Lastly, previous studies have not considered utilizing the estimated PS as a covariate in a hierarchical linear model as the sole conditioning strategy. For example, the PS was included as a

level-1 predictor variable for stratified sub-samples (Thoemmes, 2009). With this technique, the conditioning and treatment effect estimation occur simultaneously. A random coefficients model, with group membership (Z) and the estimated PS varying randomly across clusters (see Equation 24) was estimated. Z , a dummy coded variable, represented group membership and units in treatment group received a value of 1.

$$\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}Z_{ij} + \beta_{2j}PS_{ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \mu_{1j} \\
\beta_{2j} &= \gamma_{20} + \mu_{2j} \\
Y_{ij} &= \gamma_{00} + \gamma_{10}Z + \gamma_{20}PS_{ij} + \mu_{0j} + \mu_{1j}Z_{ij} + \mu_{2j}PS_{ij} + e_{ij} \\
e_{ij} &\sim N(0, \sigma^2) \\
\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \end{pmatrix} &\sim N \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \tau_{00} & & \\ & \tau_{11} & \\ & & \tau_{22} \end{pmatrix}
\end{aligned} \tag{24}$$

Under both the matching and stratification techniques there is a potential for loss of data. In this study, one control unit was matched to each treatment unit to create matched pairs, therefore, unmatched control units were dropped which resulted in a balanced sample. During this step the total number of matched pairs retained was recorded to understand the sample size of the outcome model under different conditions. The total number of matched pairs was converted into a proportion of potential matches for a given sample size condition in order to make fairly assess this conditioning method across the different design factors.

When stratifying the sample into subsamples, in this case, quintiles, each stratum needed to contain at least one unit in treatment group and one unit in control group. The absence of either renders that particular stratum futile as neither balance nor treatment effects can be estimated with units from only one group, and therefore that stratum should be dropped from the analysis; therefore, the proportion of samples creating fewer than five functional strata for each condition was reported.

Step four: Assessment of balance. One way to evaluate the estimation model and conditioning method is to examine estimates of balance between treatment and control groups. Successful estimation models and conditioning strategies remove the relationship between treatment assignment (Z) and each covariate. While some scholars attribute covariate balance to non-significant relationships, the use of null hypothesis significance testing to measure balance between groups after PS conditioning has received criticism (Austin, 2008a; Stuart, 2008). Often, and more commonly accepted, is to use a measure of effect size, such as the standardized mean difference to represent estimates of balance between groups (see equation 9). Here, the difference between the mean of the covariate for the control group is subtracted from the mean value of the covariate for the treatment group. This specific method assumes the groups are equal in size. When the groups are unequal, a pooled standard deviation, S_p , should be estimated and used as the divisor (see equation 25).

$$S_p = \sqrt{\frac{\left((n_{z_1} - 1) s_{z_1}^2 + (n_{z_0} - 1) s_{z_0}^2 \right)}{n_{z_1} + n_{z_0}}} \quad (25)$$

Here, the pooled standard deviation adjusts the sample size of each group, by subtracting the number in each group by one and multiplying by its respective variance. When assessing balance, the direction of the differences is not of substantive importance therefore often the absolute difference values are used to avoid potential bias when aggregating measures. One drawback to using standardized mean differences with MLM is that they do not account for the clustering of data. However, since conditioning was not restricted within clusters this limitation did not impose severe prohibitive implications for this study.

For each sample, the method used to assess balance and calculate standardized mean difference scores was determined based on the conditioning method. Since units were matched across clusters, the matched pair data sets were no longer nested hierarchically, but were clustered in a cross classified manner. Rather than individuals units nested in clusters only, we have

individual units nested in clusters and pairs (Raudenbush & Bryk, 2002). The following multilevel model was used to assess balance on the continuous covariates for the matched pairs.

$$\begin{aligned} X_{ijk} &= \pi_{0,jk} + \pi_{1,jk}Z_{ijk} + e_{ijk} \\ e_{ijk} &\sim N(0, \sigma^2) \end{aligned} \tag{26}$$

Here, X_{ijk} is the score on a particular covariate for the individual in cluster j and matched pair k , $\pi_{0,jk}$ is the intercept, Z_{ijk} indicates the group for the student (treatment versus control) and $\pi_{1,jk}$ is the regression coefficient relating Z_{ijk} to X_{ijk} , or the mean difference between the matched pair.

Balance for the matched pairs was conducted in SAS using the MIXED procedure for the continuous variables and the GLIMMIX procedure for the dichotomous variables.

Because conditioning across clusters was employed balance estimates on cluster level predictors was also estimated. When conditioning was restricted within clusters, balance on the cluster level predictors is fixed. With certain conditions such as large clusters with a small number of clusters, units from the same cluster may be paired. This within cluster pairing can be considered incidental. When this was the case, balance for this matched pair on the cluster predictors was considered perfectly balanced.

With regression coefficients, values closer to 0 indicate no differences; however with odds ratios, a value of 1 indicates that the event is equally likely to occur in groups. In order to consistently interpret balance for all the covariates, the log odds were rescaled to compute standardized difference scores comparable across covariates regardless of their scale of measure. Since the standard logistic distribution has a variance of $\pi^2/3$, standardized difference scores can be estimated by dividing the $\ln(\text{odds})$ by $\pi/\sqrt{3}$, or 1.81 (Chinn, 2000).

For the stratified samples, the following multilevel model was used to estimate the balance on the continuous covariates for treatment and control groups within each individual stratum. The absolute standardized differences were pooled across strata on each covariate to provide an overall estimate of balance on the entire sample.

$$\begin{aligned}
X_{ij} &= \beta_{0j} + \beta_{1j}Z_{ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \mu_{1j} \\
X_{ij} &= \gamma_{00} + \gamma_{10}Z_{ij} + \mu_{0j} + \mu_{1j}Z_{ij} + e_{ij} \\
e_{ij} &\sim N(0, \sigma^2) \\
\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} &\sim N \begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \tau_{00} & \\ & \tau_{11} \end{pmatrix}
\end{aligned} \tag{27}$$

Specifically, the γ_{10} represents the average difference on the covariates between treatment and control group within the stratum. Pooling across strata resulted in a value interpreted as the average difference on the covariate between treatment and control groups after controlling for the strata. Similarly, standardized mean differences were estimated per stratum and aggregated across stratum for each sample. As with the matched pairs, the same adjustments in SAS were made, to estimate odds ratios before they were converted to standardized difference scaled scores.

Finally, for covariate adjustment, balance was estimated for the continuous predictors using by the following model:

$$\begin{aligned}
X_{ij} &= \beta_{0j} + \beta_{1j}Z_{ij} + \beta_{2j}PS_{ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \mu_{0j} \\
\beta_{1j} &= \gamma_{10} + \mu_{1j} \\
\beta_{2j} &= \gamma_{20} + \mu_{2j} \\
X_{ij} &= \gamma_{00} + \gamma_{10}Z_{ij} + \gamma_{20}PS_{ij} + \mu_{0j} + \mu_{1j}Z_{ij} + \mu_{2j}PS_{ij} + e_{ij} \\
e_{ij} &\sim N(0, \sigma^2) \\
\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \end{pmatrix} &\sim N \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} \tau_{00} & & \\ & \tau_{11} & \\ & & \tau_{22} \end{pmatrix}
\end{aligned} \tag{28}$$

Here, γ_{10} represents the average mean difference between treatment and control groups on the covariate controlling for PS, while γ_{00} represents the average mean on the covariate for the control group controlling for PS. Standardized differences were estimated on the full sample. Again, adjustments and conversions were made to estimate balance for the dichotomous predictors.

For this study, a covariate was considered imbalanced when differences between groups were larger than 0.25. This criterion may be considered liberal however, because previous studies focused on conditioning within clusters, there is no empirical evidence to suggest what constitutes a reasonable criterion in this context. Additionally, conditioning across clusters introduces cluster variability accordingly setting a smaller criterion does not seem pragmatic. Literature suggests entire samples be considered imbalanced if more than 10% of the covariates are imbalanced (Ho, Imai, & King, & Stuart, 2007; Rubin, 2001, Shadish & Steiner, 2010). Using this criterion, an entire sample would be classified as unbalanced if standardized differences greater than .25 are present in 5 or more covariates. In addition to this nominal classification an estimate of the degree of balance will be computed across all covariates for each sample and aggregated across replications.

Step five: Estimate treatment effect. In order to further assess the quality of the estimation methods and conditioning strategies different outcome models corresponding to the different conditioning strategies were used to estimate the treatment effects and provide information regarding the bias, RMSE, 95% confidence interval coverage, 95% confidence interval width. The outcome models used to estimate the treatment effects for matching and stratification were identical to those used to estimate balance on the continuous covariates, replacing the continuous covariate X , for the continuous outcome Y (see equation 26 for matching and equation 27 for stratification). For covariance adjustment, the outcome model is the conditioning model and is represented by equation 24.

Bias was calculated as the average difference between the known parameter and the estimated parameter value across all replications for each condition sampled and reported across the 12 PS approaches. The RMSE of the estimated treatment effects was calculated for the entire set of N replications for each condition.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{1000} (\hat{\delta} - \delta)^2}{1000}} \quad (29)$$

The confidence interval width was calculated as the difference between the upper and lower limit of the 95% confidence interval around the estimated treatment effect for each replication and aggregated over the entire set of replications per condition. The confidence interval coverage was defined as whether a 95% confidence interval included the true treatment effect for each replication and a coverage proportion will be calculated for each condition.

Step six: Final comparative analysis. To investigate the performance of the different PS methods, defined by estimates of balance and treatment effects, and answer the research questions a factorial ANOVA including all the design factors was computed for each dependent variable, where the PS estimation methods and PS conditioning techniques were considered within-subjects factors and the sample characteristics represented between-subjects factors.

The results of the simulation were analyzed using PROC GLM in SAS 9.3 for both the balance and treatment effect estimates such that the dependent variables were balance score, number of unbalanced covariates, proportion of samples balanced, bias, RMSE, confidence interval coverage, and confidence interval width and the independent variables were the two PS method factors and the seven data factors. Models were built with the purpose of finding effects whose eta-squared values were .0588 or greater, which according to Cohen's (1988) standards would be considered a moderate effect. The effect size, or eta-squared (η^2) values were calculated to determine the proportion of variability associated with each effect. Each model was created using all main effects and first order interactions. If this model explained more than 90% of the total variability, then no further models were investigated. However, if less than 90% of the total variability was explained then all three way and at times four way interactions were included in the models. These models were compared with the original models to see if any additional interaction effects explained at least 5.8% of the variability, or had eta-squared values

of at least 0.588. If no additional interactions were significant, then the simplest model with significant interactions was interpreted. Finally, to investigate the relationship between balance and treatment effects, and answer research question 5, a series of correlations were computed between the estimates of balance and treatment effects.

Chapter Summary

This chapter outlined the proposed methodology for this study and describing the purpose, research questions, design, sample characteristics, data generation methods, analytical procedures, and outcome measures. One goal of this chapter was to illustrate the lack of empirical evidence available to guide methodologists let alone applied researchers seeking to use propensity scores as a method to adjust for selection bias in educational or other multilevel contexts, to ultimately justify the need for not only this current study but for many investigations within this framework.

CHAPTER FOUR: RESULTS

This chapter presents the results of the study organized in order of analytical procedures. First, the purpose and research questions are presented followed by a description of the samples after PS estimation, trimming and conditioning. Next, the analytical procedures are described and organized by outcome variable. Finally, the results from the analytical procedures are used to answer the study's research questions.

Overview of the Study

The purpose of this study was to examine appropriateness of using PS methods to achieve balance between groups on observed covariates and to yield unbiased treatment effect estimates in multilevel studies. Specifically, this study examined the extent to which different PS approaches (PS estimation models and PS conditioning techniques) and sample characteristics (sample size, covariate relationship to treatment and outcome, and population effect size) achieved balance and reproduced the population treatment effect. To meet this goal, five different research questions were posed:

1. To what extent do balance estimates vary across PS methods (PS estimation models and PS conditioning strategies)?
2. To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the balance achieved by the PS methods (PS estimation models and PS conditioning strategies)?
3. To what extent do treatment effect estimates vary across PS methods (PS estimation models and PS conditioning strategies)?

4. To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the treatment effects estimated by the PS methods (PS estimation models and PS conditioning strategies)?
5. What is the direction and strength of the relationship between balance and both the accuracy and precision of treatment effect estimates?

Description of Samples

Propensity Score Estimation Models

For each sample, four different PS models, (one single level and three multilevel) were estimated and examined. In order to compare the PS distributions across the models the mean PS, variance of the PS in the sample, and the inter-quartile range for the treatment and control groups were calculated for each sample and aggregated across all replications and presented in Table 5. The single level logistic model and the random intercepts model produced nearly identical PS distributions for treatment and control groups. Similarly, the PS distributions for treatment and control groups were nearly identical for the random coefficients model and the cross level model.

Table 5
Average Distributions of the Propensity Scores for each Estimation Model for Treatment and Control Groups

PS Model	Mean		Variance		Inter-quartile range	
	Treatment	Control	Treatment	Control	Treatment	Control
SLL	.70	.21	.04	.05	.37	.27
RI	.70	.20	.04	.05	.37	.26
RC	.75	.17	.03	.04	.31	.22
CL	.75	.17	.03	.03	.31	.22

Next, for each sample generated, correlations between the estimated PSs were computed using Pearson product-moment correlation coefficient estimates and averaged across replications. The data show a strong positive relationship across the models. The strength of the relationship between the single level model and the multi-level models decreases as the model becomes more complex. Mean correlations are presented in Table 6.

The distribution of non-convergence rates for each estimation model is illustrated in Figure 3. The average proportion of samples not converging across all models was close to zero, with average rates increasing with model complexity. The single level model had an average non-convergence rate of .03 ($SD=.09$) and a range of average rates from 0.0 to .48. The random intercepts and random coefficients model had similar non-convergence rates with the average values of .07 ($SD=.20$) and .08 ($SD=.21$) respectively. The range of mean non-convergence values was larger for the multilevel models as compared to the single level model. The random intercepts model had a range of mean rates from 0.0 to .88 and the random coefficients model had a range of mean rates from 0.0 to .92. Finally the cross-level model had an average non-convergence rate of .10 ($SD=.25$) and a range of average rates from 0.0 to .97.

Table 6
Descriptive Statistics for Mean Correlations between Propensity Score Estimates

PS Models	Mean	Sd	Min	Max
corr(SLL,RI)	.985	.008	.944	1.0
corr(SLL,RC)	.948	.028	.849	.993
corr(SLL,CL)	.938	.032	.756	.992
corr(RI,RC)	.967	.018	.869	.988
corr(RI,CL)	.957	.028	.647	.998
corr(RC,CL)	.993	.004	.942	.999

The variation in mean non-convergence rates was explored by modeling the proportions with eight main effects (number of clusters, within cluster sample size, relationship between level-1 covariates and treatment assignment, relationship between level-1 covariates and outcome, relationship between level-2 covariates and treatment assignment, relationship between level-2 covariates and outcome, population effect size, and PS estimation method), and all possible two-way interactions. The eta squared values (η^2) values were examined, and those values exceeding Cohen's (1988) medium effect size criteria of $\eta^2 = .0588$ or greater were considered significant and were further explored through box plots. This criterion was used for this model and all subsequent models in order to explain any differences in variables as a function of the PS models, conditioning methods and sample characteristics.

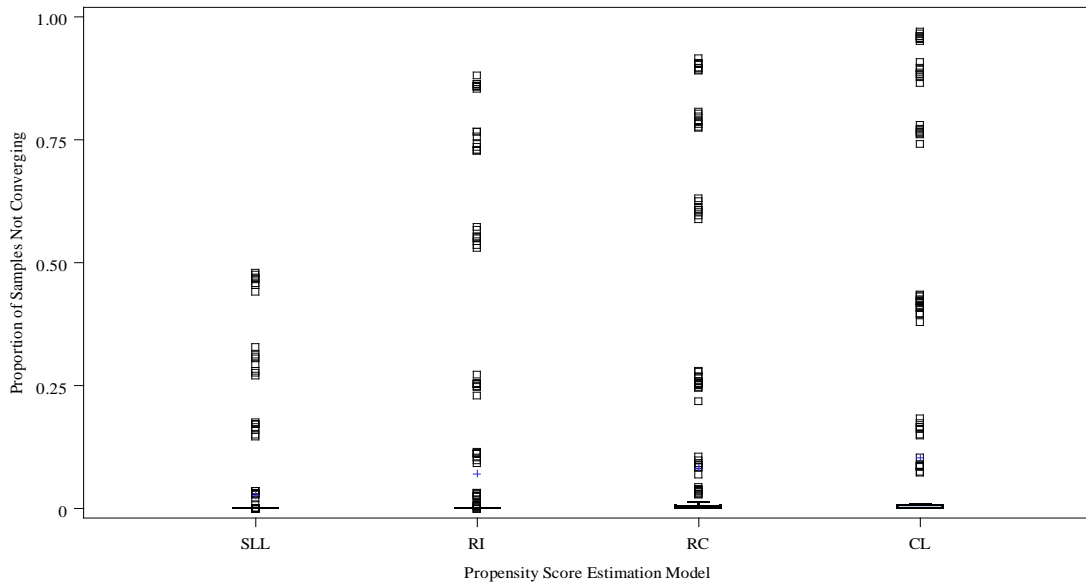
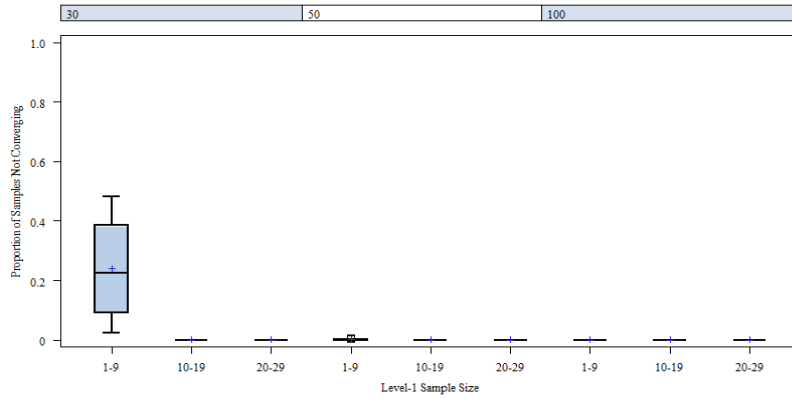


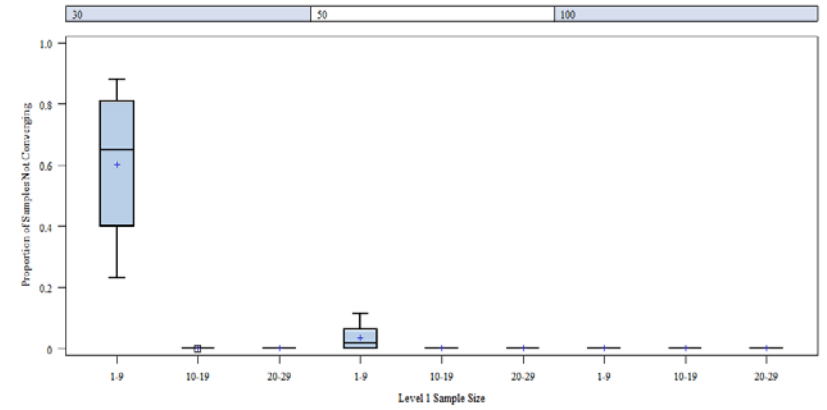
Figure 3. Mean non-convergence distributions by propensity score estimation model

The model with main effects and all possible two-way interactions explained more than 94% of the variability in mean non-convergence rates. Both the level 1 sample size ($\eta^2=.24$) and the level 2 sample size ($\eta^2=.16$) emerged as major factors related to non-convergence. The interaction between the level 1 and level 2 sample size factors was also associated with the variability in non-convergence rates ($\eta^2=.33$). On average, more than half of the samples with few level 1 units within a cluster (1-9) and a small number of clusters (30) did not converge. Specifically, the mean non-convergence rate for this interaction level was .56 ($SD=.29$). When increasing the number of clusters from 30 to 50, with few level 1 units, the mean non-convergence rate decreased to .07 ($SD=.11$). All other levels of the interaction had an average non-convergence rate less than .003. The distributions of the mean non-convergence rates for this interaction are presented in Figure 4 by PS estimation model. Noteworthy in this figure is the increasing pattern of non-convergence rates for the two interaction levels (1-9 and 30; 1-9 and 50) as the PS model grew in complexity.

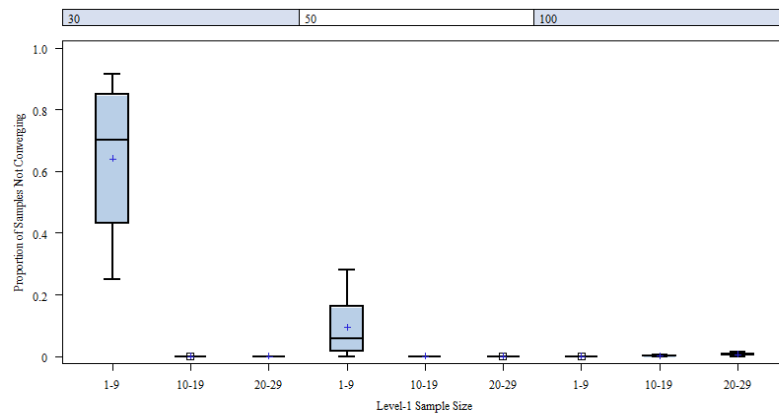
Single Level Logistic Model



Random Intercepts Model



Random Coefficients Model



Cross-Level Model

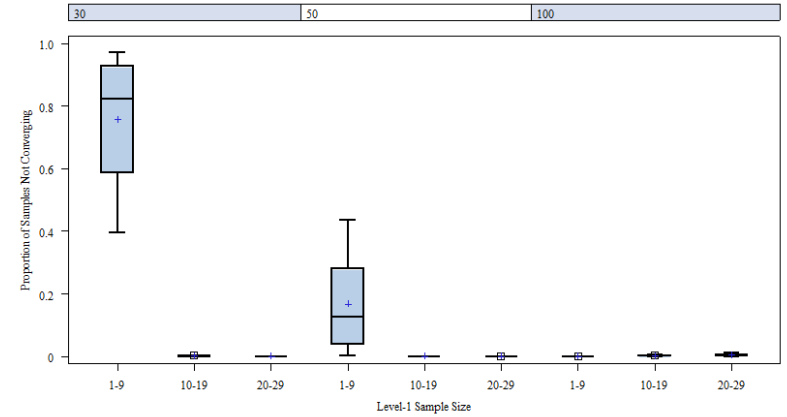


Figure 4. Mean non-convergence distributions by level 1 sample size across the number of clusters for each PS model

The distribution of non-positive definite matrix rates for each estimation model is illustrated in Figure 5. Estimating non-positive definite matrices is an issue that may occur with multi-level models. The random intercepts model had a mean non-positive definite matrix rate of .05 ($SD=.05$) and a range of mean rates from 0.0 to .27. The random coefficients model had the largest mean non-positive definite matrix rate of .71 ($SD=.26$) and a range of mean rates from .12 to .985. Finally, the cross-level model had a mean non-positive definite matrix rate of .70 ($SD=.26$) and a range of mean rates from .12 to .995.

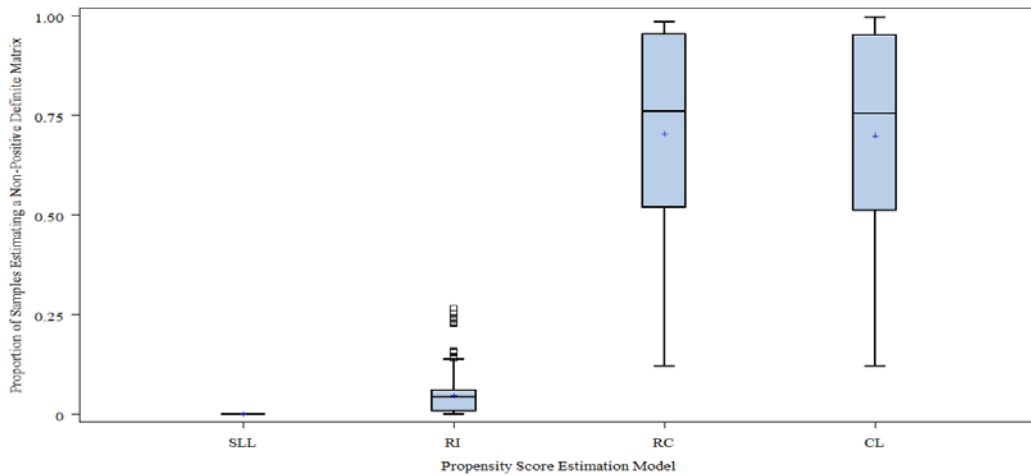


Figure 5. Mean non-positive definite matrix rates by propensity score estimation model

The variation in non-positive definite matrix rates was further explored using the same model described above. The mean proportions were modeled with the eight main effects and all first order interactions. The model accounted for more than 98% of the variation in the non-positive definite matrix rates. Factors related to the non-positive definite matrix rates included estimation model ($\eta^2=.77$), level 1 sample size ($\eta^2=.09$) and the interaction between estimation model and the level 1 sample size ($\eta^2=.07$). The mean non-positive definite matrix rates associated with this interaction are presented in Table 7. The random coefficients and cross-level models had high mean rates of non-positive definite matrices across all levels of the level

1sample size factor. As the sample size level increased the mean proportion samples estimating non-positive definite matrices decreased.

Table 7
Descriptive Statistics for Mean Non-Positive Definite Matrix Rates for each Multilevel PS Estimation Model by Level 1 Sample Size

PS Model	Level 1 Sample Size											
	1-9				10-19				20-29			
	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max
RI	.08	.06	.03	.26	.04	.03	0	.11	.017	.02	0	.07
RC	.96	.02	.91	.99	.72	.14	.40	.91	.43	.19	.12	.75
CL	.96	.019	.91	.99	.71	.14	.39	.90	.43	.19	.12	.74

Common Support

After the propensity scores were estimated, the samples were trimmed to retain only the region of common support for the treatment and control units. Table 8 compares the average extreme values for the treatment and control groups before trimming against the distributions of the trimmed sample for each PS estimation model. After trimming, the range of the PSs was still quite large, getting marginally smaller as the model grew in complexity.

Table 8
Mean Propensity Score Range Before and After Trimming

PS Model	Before				After	
	Control		Treatment		Retained Sample	
	Low	High	Low	High	Low	High
SLL	.001	.93	.051	.995	.051	.93
RI	.0009	.92	.047	.995	.05	.92
RC	.0005	.86	.10	.997	.10	.84
CL	.0006	.86	.10	.998	.11	.83

The distribution of the mean percent of data trimmed by PS estimation model is illustrated in Figure 6. As the model became more complex, the mean percent of non-overlapping data increased. The single level logistic model had an overall mean percent of data trimmed of 20.55 ($SD= 15.64$) with mean percents ranging from 2.44 to 96. The random intercepts model had an overall mean percent of data trimmed of 24.11 ($SD= 18.42$) ranging from

3.03 to 100. The random coefficients and cross-level models had nearly identical distributions with an overall mean percent of data trimmed of 37.21 ($SD= 24.25$) and 37.75 ($SD= 23.83$) respectively. For the random coefficients model the mean percent of data trimmed ranged from 9.02 to 100, and for the cross level model the mean percents ranged from 0 to 100.

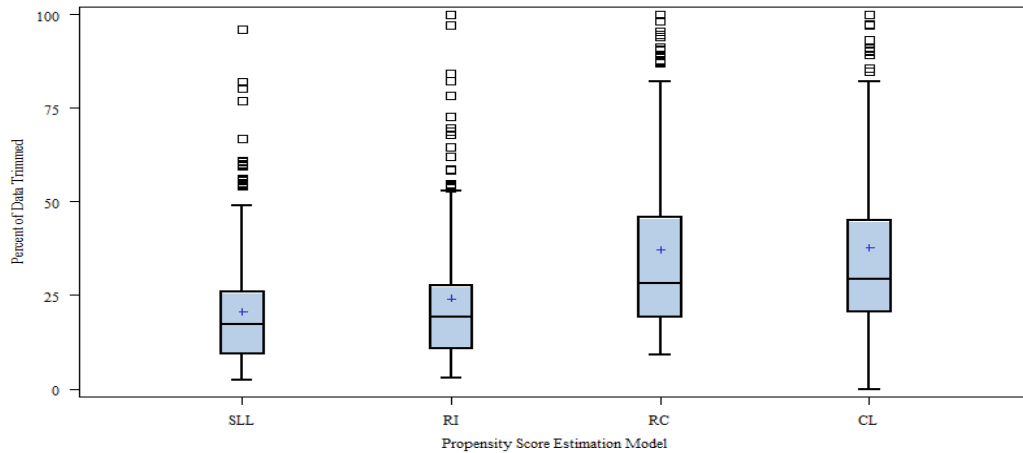
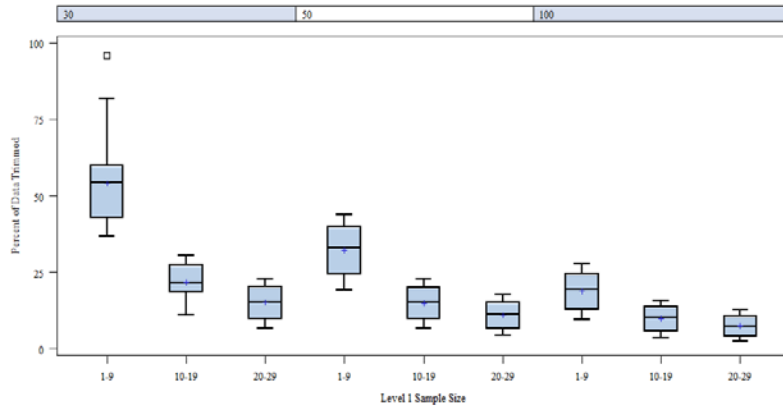


Figure 6. Mean percent of data trimmed by propensity score estimation model

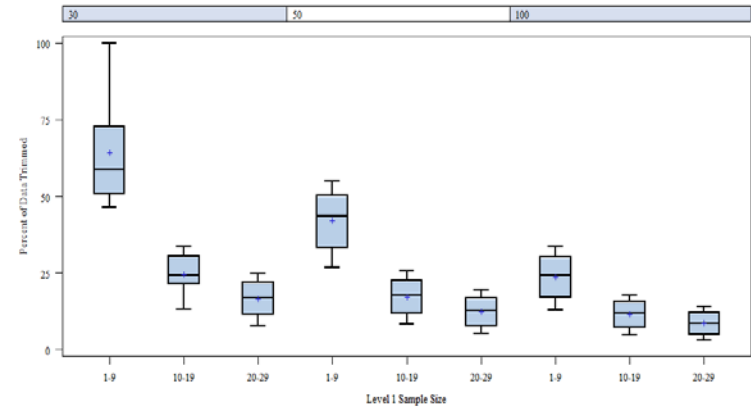
The variation in mean percent of data trimmed was further explored across the eight main effects and all first order interactions. The model explained nearly 96% of the variability in the mean percent of data trimmed. Several factors impacted the mean percent of data trimmed: level 1 sample size ($\eta^2=.43$), level 2 sample size ($\eta^2=.18$), estimation model ($\eta^2=.14$), the interaction between level 1 sample size and level 2 sample size ($\eta^2=.07$), and the level 1 covariate relationship to treatment ($\eta^2=.06$). Figure 7 displays the interaction between level-1 sample size and level-2 sample size for each PS Model. As the level-1 sample size increased the percent of data trimmed decreased. This pattern was consistent across the different cluster values for each PS model. A large percent of data was trimmed for the conditions with 1-9 units within 30 clusters and grew as the models became more complex (see Figure 7).

Finally, the level-1 covariate relationship to treatment also impacted the percent of data trimmed. As the strength of the relationship between the level-1 covariates and treatment

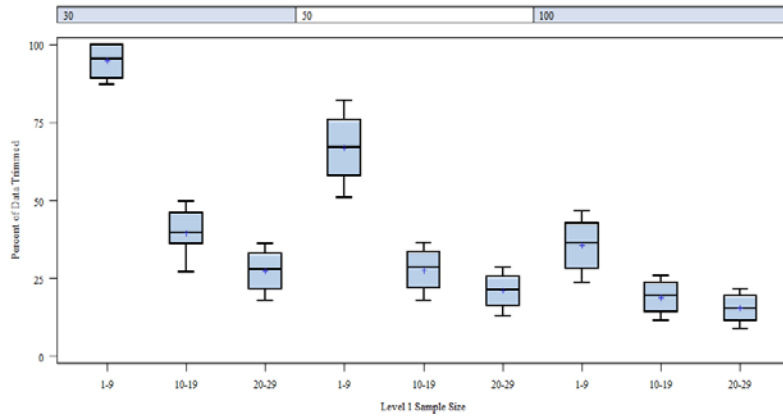
Single Level Logistic Model



Random Intercepts Model



Random Coefficients Model



Cross-Level Model

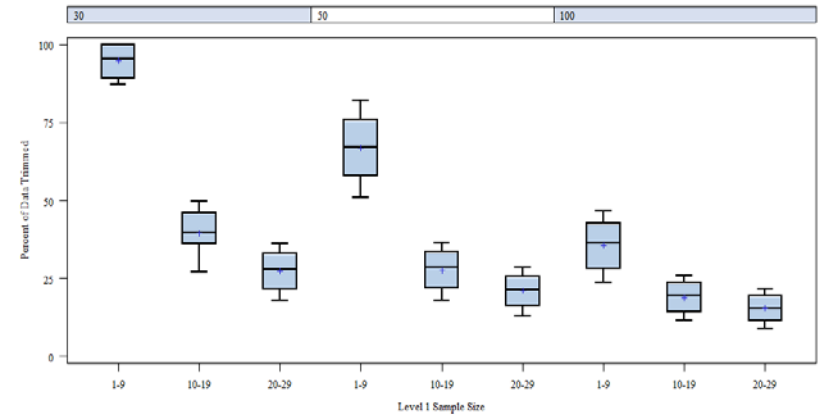


Figure 7. Mean Percent of Data Trimmed by level-1 sample size across level-2 sample size for each propensity score model

assignment increases the greater percentage of data were trimmed. When the level-1 covariate relationship to treatment was small ($\beta_{xz}=.10$), the overall mean percent of data trimmed was 25.11($SD=22.05$) with mean percents ranging from 2.4 to 100. When the level-1 covariate relationship to treatment was moderate ($\beta_{xz}=.20$), the overall average mean increased 34.64($SD=21.22$) with mean percents ranging from 0 to 100. Table 9 explores these means across PS Model and the same pattern holds true across all models with the percent in data trimmed increasing with the complexity in the model.

Table 9

Mean Percent of Data Trimmed for Covariate Relationship to Treatment by PS Model

Parameter	PS Model				
	SLL	RI	RC	CL	
β_{xz}	.10	14.93	18.49	32.59	34.53
	.20	26.17	29.85	41.99	41.17

Note. β_{xz} = strength of relationship between level-1 covariates and treatment assignment

Propensity Score Conditioning

After the samples were trimmed three different PS conditioning techniques were conducted (covariance adjustment, matching, and stratification). Both matching and stratification methods may result in the loss of data. When matching units, a caliper width of .25 between the units estimated PS was imposed. If a control unit whose estimated PS was not within .25 of a treatment unit it was not matched. Likewise, if there were not adequate matches for treatment units it was not matched. All unmatched cases were dropped from the sample. When stratifying the samples, strata that did not have at least 2 treatment units and 2 control units were considered incomplete and were dropped from the analysis. To investigate if certain design factors impacted non-matched cases and empty strata the mean proportion of potential matches and the proportion of samples dropping strata were explored.

The mean proportion of potential matches was calculated by taking the ratio of the mean number of matches and the mean number of treatment units in the sample post trimming. At times, this proportion was larger than 1.0 when the mean number of matches was larger than the

mean number of units in treatment group post trimming. This occurred when level-1 sample size was 1-9 and the number of clusters was 30 across all four PS estimation models. For example, for one of these conditions the mean number of treatment units was 2.00 and for the same condition the mean number of matched cases was 3.91, resulting in a mean proportion of matched cases of 1.96. Additionally, since the number of units (1-9) was uniformly distributed and there was a 1:3 treatment to control ratio, samples may have had specific clusters with no treatment units or very few treatment units which may have been dropped after trimming; thus impacting the mean number of treatment units post trimming. Samples with proportions greater than 1 were removed from this analysis only.

Figure 8 displays the distributions of the mean proportion of potential matches by PS Model. In general, the typical value for the mean proportion of potential matches was comparable across PS estimation models dropping slightly as the model complexity increased. The single level logistic model had an overall mean proportion of potential matches of .62 ($SD=.09$) with values ranging from .50 to .80. The mean proportions of potential matches dropped slightly as compared to the single level model for the random intercepts and random coefficients model, but the range of values got larger. The random intercepts model had an overall mean proportion of potential matches of .60 ($SD=.10$) with values ranging from .31 to .90. Similarly, the random coefficients model had an overall mean proportion of potential matches of .60 ($SD=.10$) with values ranging from .01 to .97. Samples with a .97 mean proportion of potential matches suggests that on average almost all of the treatment units remaining in the sample after trimming were matched. The range in values decreased for the cross level model which had an overall mean proportion of potential matches of .54 ($SD=.10$) ranging from 0 to .67. A mean proportion of potential matches of 0 indicate on average none of the treatment units in the sample after trimming were matched.

Variation in the mean proportion of potential matches was explored across the eight main effects and all first order interactions. This model explained 84% of the variability in the mean

proportion of potential matches. Factors impacting the mean proportion of potential matches included the level-1 covariate relationship to treatment ($\eta^2=.47$); the level-2 covariate relationship to treatment ($\eta^2=.16$), and the estimation model ($\eta^2=.08$).

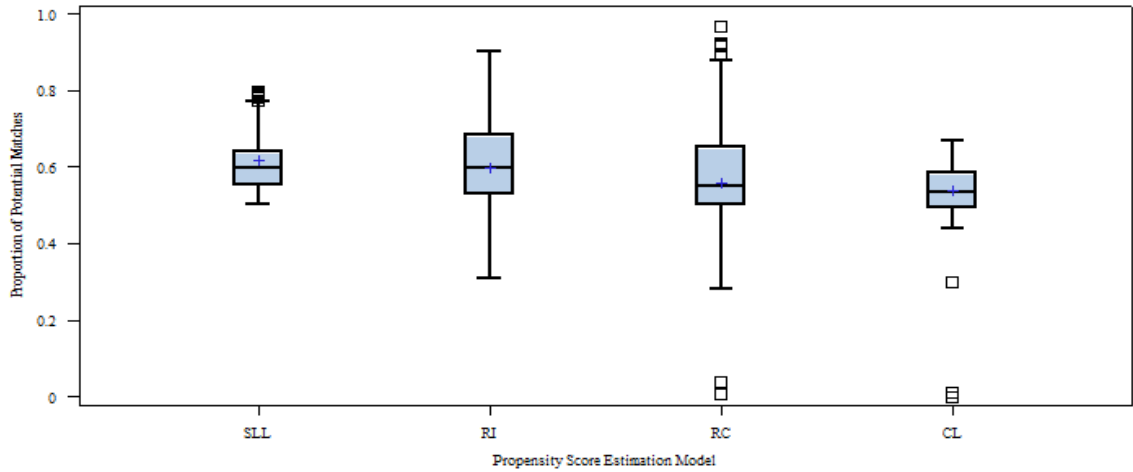


Figure 8. Distributions of mean proportion of potential matches across PS models

When the level-1 covariate relationship to treatment was small ($\beta_{xz}=.10$), the overall mean proportion of potential matches was $.65(SD=.08)$ with mean proportions ranging from $.53$ to $.97$. When the level-1 covariate relationship to treatment was moderate ($\beta_{xz}=.20$), the overall mean proportion of potential matches was $.51(SD=.06)$ with mean proportions ranging from 0 to $.59$. When the level-2 covariate relationship to treatment was small ($\gamma_{0w}=.20$), the overall mean proportion of potential matches was $.62(SD=.11)$ with mean proportions ranging from $.0$ to $.97$. When the level-2 covariate relationship to treatment was moderate ($\gamma_{0w}=.40$), the overall mean proportion of potential matches was $.54(SD=.07)$ with mean proportions ranging from 0 to $.64$. Mean proportions of potential matches for both the level-1 covariate relationship to treatment and the level-2 covariate relationship to treatment by PS model are presented in Table 10. The proportion of potential matches was greater when the strength of the relationship between covariates and assignment was small. This pattern was consistent for both the level-1 and level-2

covariates. The proportion of matches decreased marginally as the PS model became more complex.

Table 10
Mean Proportion of Potential Matches for Covariate Relationship to Treatment by PS Model

Parameter	PS Model			
	SLL	RI	RC	CL
β_{xz}	.10	.70	.68	.63
	.20	.54	.52	.49
γ_{0w}	.20	.66	.65	.60
	.40	.58	.55	.51

Note. β_{xz} = strength of relationship between level-1 covariates and treatment assignment
 γ_{0w} = strength of relationship between level-2 covariates and treatment assignment

When stratifying the samples, strata that did not have at least 2 treatment units and 2 control units were considered incomplete and were dropped from the analysis. The distributions of the proportion of samples dropping at least one stratum by propensity score model are presented in Figure 9. The average proportion of samples dropping strata was similar for the single level logistic ($M=.10, SD=.21$) and random intercepts ($M=.08, SD=.15$) model. Likewise, the random coefficients and cross-level models performed similarly. The average proportion of samples dropping strata for the random coefficients model was .16 ($SD=.23$) and was .15 ($SD=.22$) for the cross-level model.

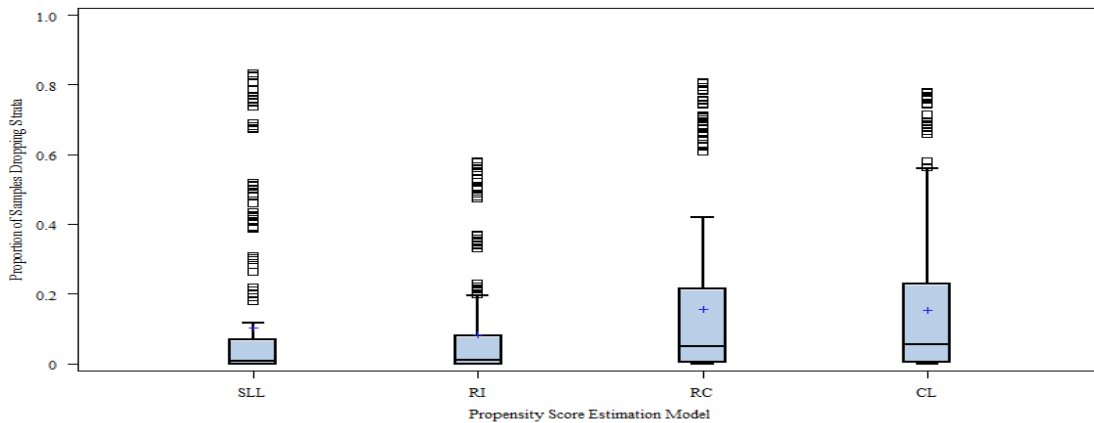


Figure 9. Proportion of samples dropping strata by propensity score estimation model

Variation in the proportion of samples dropping strata was explored across the eight main effects and all first-order interactions. This model explained 60% of the variability in the proportion of samples dropping strata. Variation was further explored across the eight main effects, all first-order interactions, and all second-order interactions. This model explained 84% of the variability in the proportion of samples dropping strata. Factors impacting the proportion of samples dropping strata included level-1 sample size ($\eta^2=.31$), the second-order interaction of the level-1 sample size, the level-2 sample size and the PS estimation model ($\eta^2=.14$), the first-order interaction between the level-1 sample size and the level-2 sample size ($\eta^2=.12$) and the level 2 sample size (.08). Figure 10 illustrates the second-order interaction effect displaying the PS estimation model by level-1 sample size interaction across the three cluster levels. A larger proportion of strata were dropped when the level 1 sample size was small and as the model grew in complexity. These patterns were fairly consistent across the cluster levels.

Data Analysis

This study intended to be a fully crossed and balanced factorial design with seven between-subjects factors fully crossed with two within-subjects factors (i.e. propensity score estimation models and propensity score conditioning methods) resulting in a total of 3,456 conditions. Due to the PS estimation method non-convergence rates, the trimming of the samples, the dropping of cases during matching and stratification, and lastly the non-convergence of the final treatment effect models, there were quite a few conditions when the level 1 sample size was 1-9 and the level-2 sample size was 30, with missing outcome variables. In a balanced design, if convergence and loss of data were not an issue, each of the 9 sample size levels (level 1 crossed with level 2) would have 384 observations, as there would be 32 unique levels per sample size interaction level and 12 observations per unique factor (four PS models crossed with the three conditioning methods). Given the convergence issues, and the data loss attributed to trimming and PS conditioning, the smallest sample size interaction level, where there were 1-9 level 1 units and 30 clusters, returned 100 observations on the outcome variables of interest.

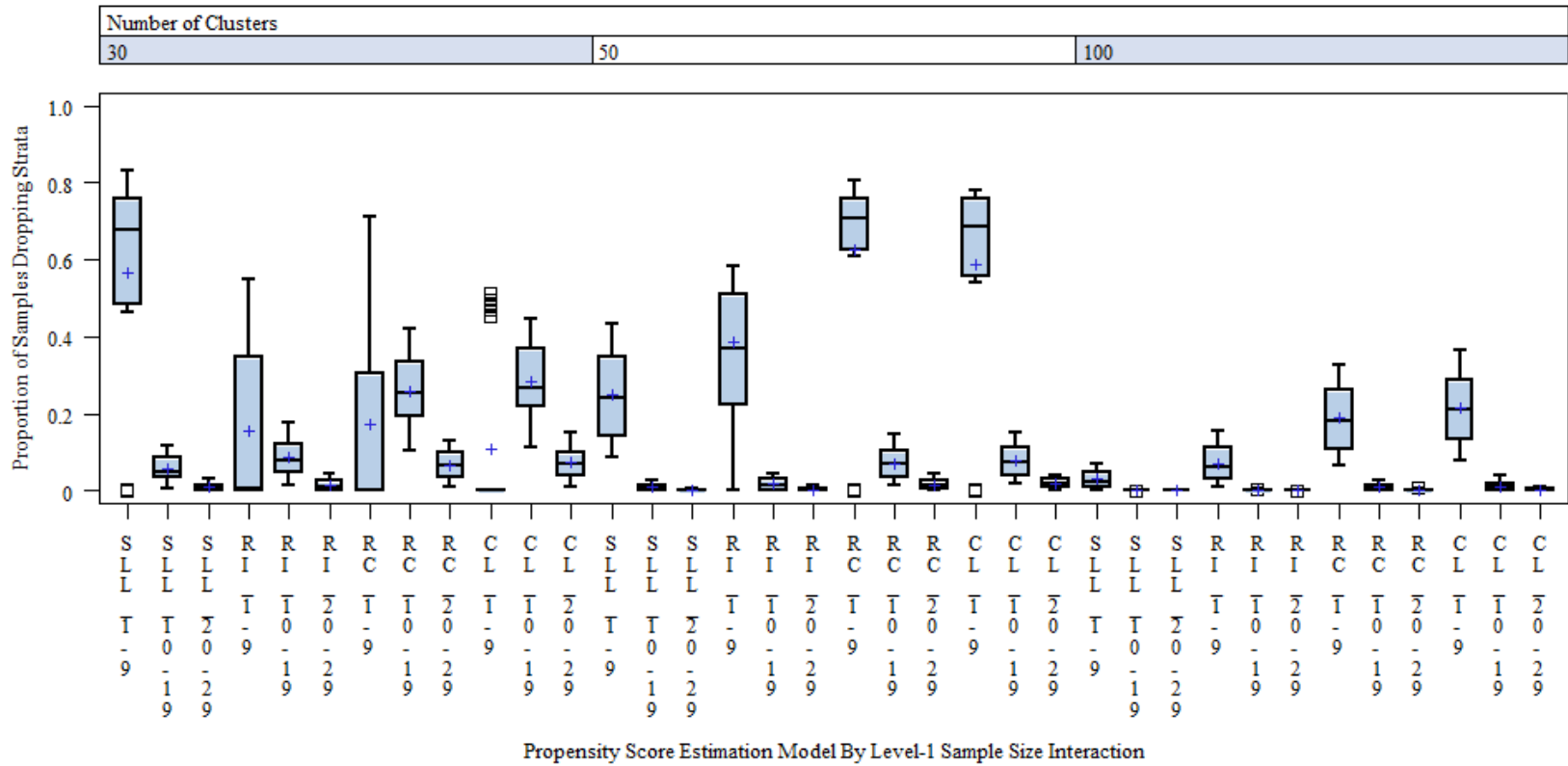


Figure 10. Proportion of Samples Dropping Strata by the Sample Size Interaction across the Four Propensity Score Models

Each of the remaining 8 levels had all 384 observations. What follows below, is a discussion on the three outcome measures associated with balance followed by a discussion on the four outcome measures associated with the treatment effect estimates.

Balance

Three different outcome measures were used to operationalize balance in this study. The first, the absolute value of the standardized mean difference between groups on the covariates, was used as the overall balance score, with smaller standardized mean differences indicating the samples were approaching a state of balance. There are no general guidelines that suggest what the cut off score should be to indicate the differences were so small that the groups were balanced. Given that 0.25 was used as the matching caliper in this study and that this was one of the seminal studies assessing PS balance using MLM and that there were both continuous and dichotomous covariates included 0.25 was used as the general guideline for what constituted acceptable balance. Although some may consider this a liberal index, it is still within the acceptable range. Additionally, with the three dichotomous covariates at level-1 and one at level-2, as well as the conversion of their log odd estimates to standardized mean differences, this index is fair.

The absolute value of the standardized mean difference between treatment and control groups for each variable was calculated and combined to provide the sample with an overall balance score. Figure 11 displays the distributions of the mean balance score for each PS factor (conditioning methods and estimation models). Notable in this figure is the overall poor performance of stratification across all PS models, with typical values approaching the unbalanced index of .25 for the single level logistic and random intercepts models, and greater than .25 for the random coefficients and cross-level models. In contrast, covariance adjustment and matching provide adequate balance estimates across all PS models. As the models became more complex the typical balance score increased for all the conditioning methods. Covariance adjustment performed slightly better under the single level logistic and random intercepts model

while matching performed slightly better as the model became complex, specifically the random coefficients and the cross level model.

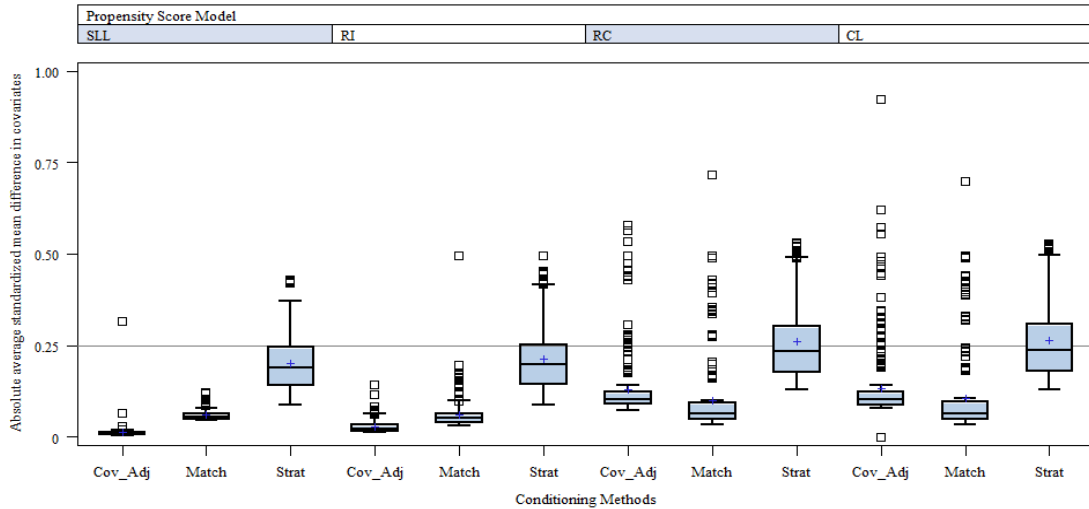


Figure 11. Absolute average standardized mean difference in covariates by conditioning method across the four PS models

Variation in the mean balance score was explored across the nine main effects and all first-order interactions. This model explained 92% of the variability in the average balance score. Main effects for conditioning method ($\eta^2=.40$), level-1 sample size ($\eta^2=.19$), estimation method ($\eta^2=.09$), and level-2 sample size were ($\eta^2=.07$) were all associated with the variability in mean balance score. Descriptive statistics for the mean balance estimates for these related factors are presented in Table 11. Overall, stratification on average performed poorly. Additionally, when level-1 sample size was 1-9, the typical mean balance score and range were both quite large. It is not clear to what extent the mean balance score when level-1 sample size was 1-9 would be comparable to the other two sample size levels (10-19 and 20-29) had all the samples converged. What is interesting to note are the range of values for some design factors. The typical values for the different PS models were comparable, but the range of mean balance scores increased tremendously as the models became more complex. The cross-level model produced mean

balance values as high as .93. Finally, as the number of clusters increased the typical mean balance score as well as range of values decreased.

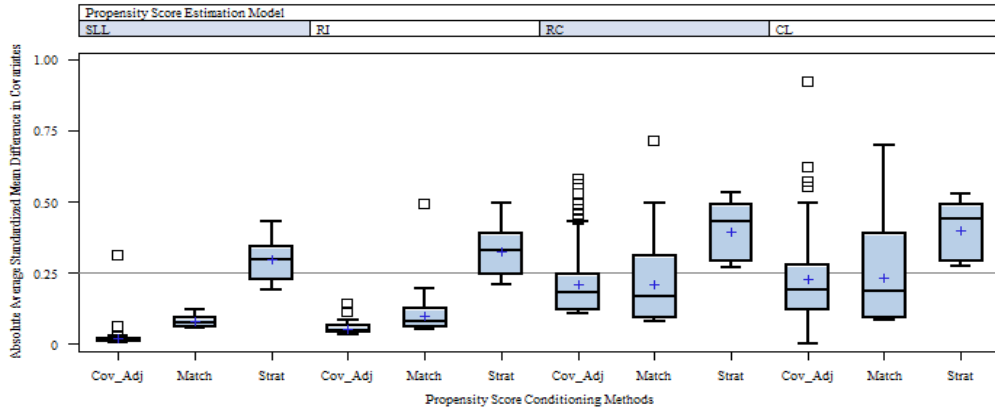
Table 11
Descriptive Statistics by Design Factors Associated with Mean Balance Score

Design Factor	Mean	SD	Min	Max
Conditioning Method				
Covariance adjustment	0.07	0.09	0	0.93
Matching	0.08	0.08	0.03	.72
Stratification	0.23	.10	.08	.53
Level-1 Sample Size				
1-9	.21	.16	0	.93
10-19	.11	.86	.005	.34
20-29	.09	.06	.004	.25
PS Estimation Model				
SLL	.09	.09	.004	.43
RI	.10	.09	.014	.50
RC	.16	.12	.04	.72
CL	.17	.12	0	.92
Level-2 Sample Size				
30	.15	.13	0	.93
50	.15	.13	.006	.53
100	.09	.08	.004	.32

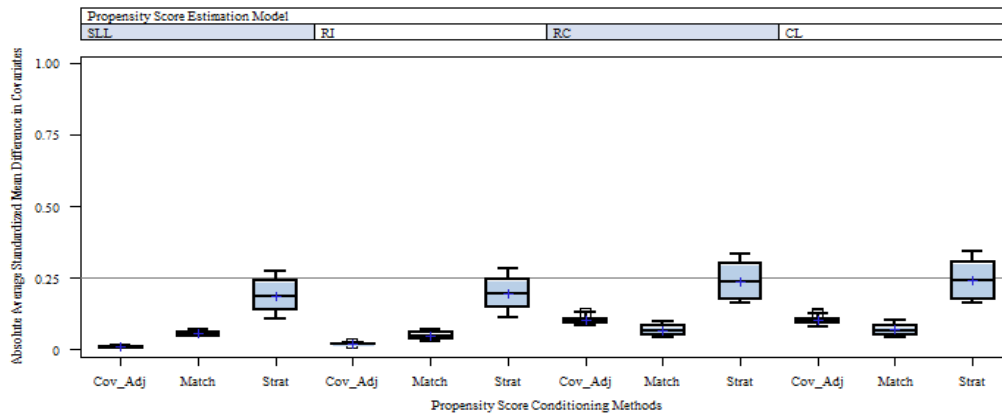
Figure 12 displays the distributions of the level-1 sample size main effect across all PS factors (conditioning methods and estimation models) and Figure 13 displays the distributions of the level-2 sample size main effect across all PS factors.

The second outcome measure used to evaluate balance in the study was the mean number of variables not balanced. For each variable, the absolute standardized mean difference score between treatment and control groups was calculated. The number of covariates with absolute standardized mean difference scores greater than 0.25 was counted. Figure 14 displays the mean number of unbalanced covariates for each conditioning method across the four PS models. A sample was considered unbalanced if more than 10% of the covariates (five covariates) had absolute standardized mean difference scores greater than .25. Evident in this figure is the overall poor performance of stratification. Across all PS models, the typical mean number of unbalanced

Level -1 Sample Size =1-9



Level -1 Sample Size =10-19



Level -1 Sample Size =20-29

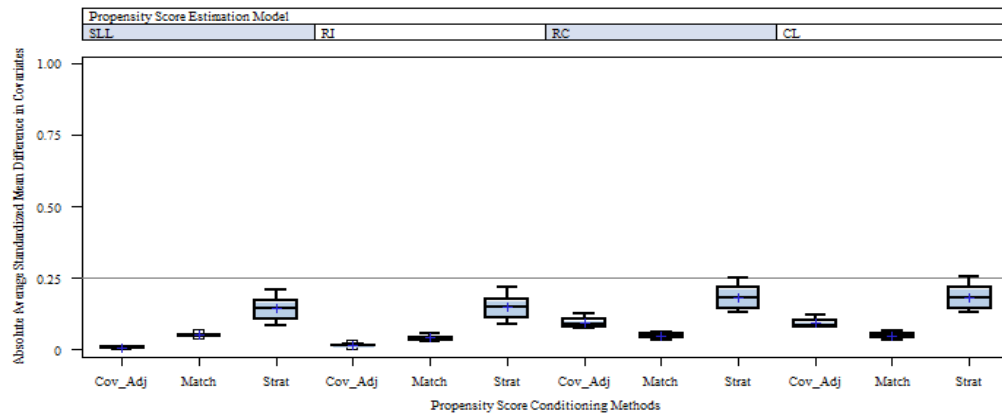
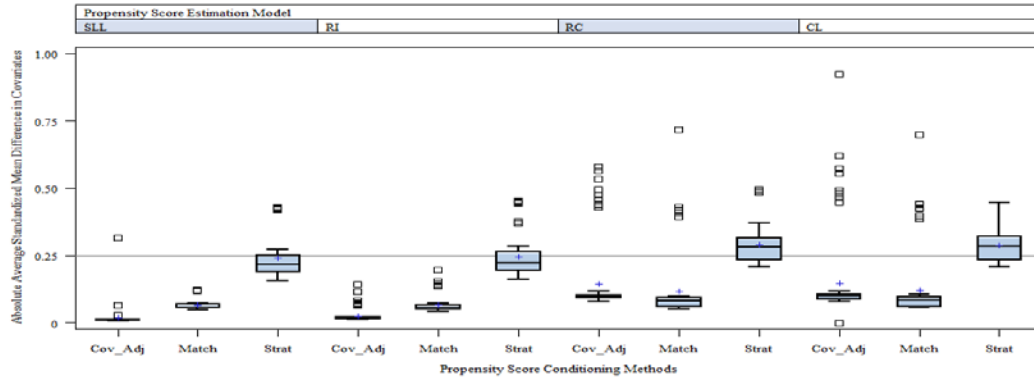
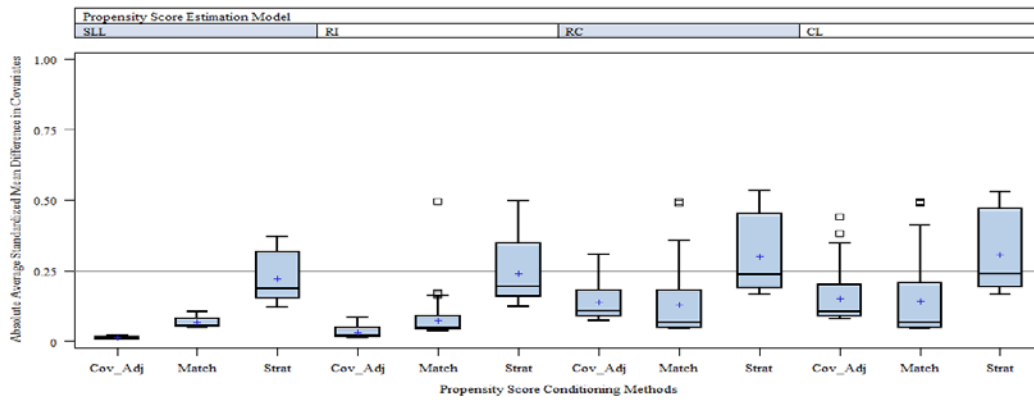


Figure 12. Distributions of the absolute average standardized mean difference in covariates by conditioning method across the four PS models for each level-1 sample size level

Number of Clusters=30



Number of Clusters=50



Number of Clusters=100

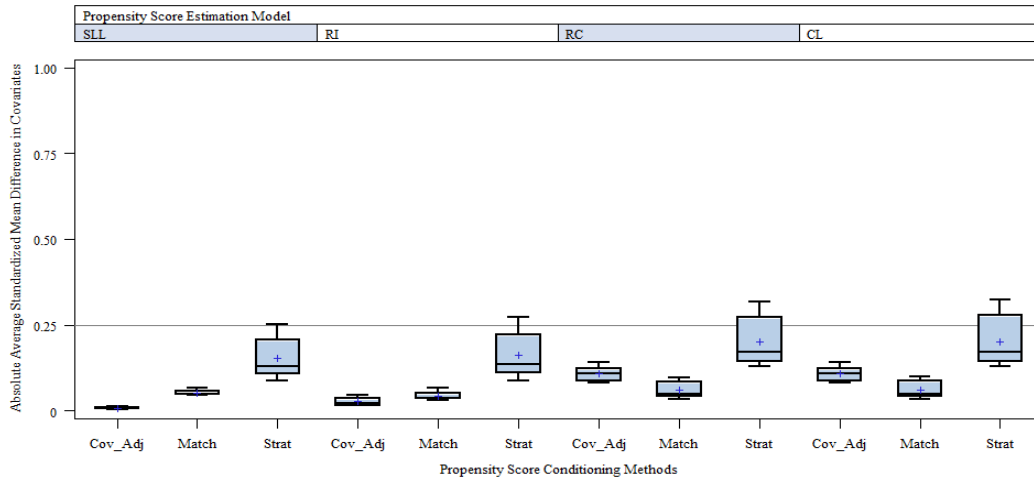


Figure 13. Absolute average standardized mean difference in covariates by conditioning method across the four PS models for each level-2 sample size level

covariates with stratification was larger than the 10% cut off. In contrast, the median values for covariance adjustment and matching were all below the cutoff.

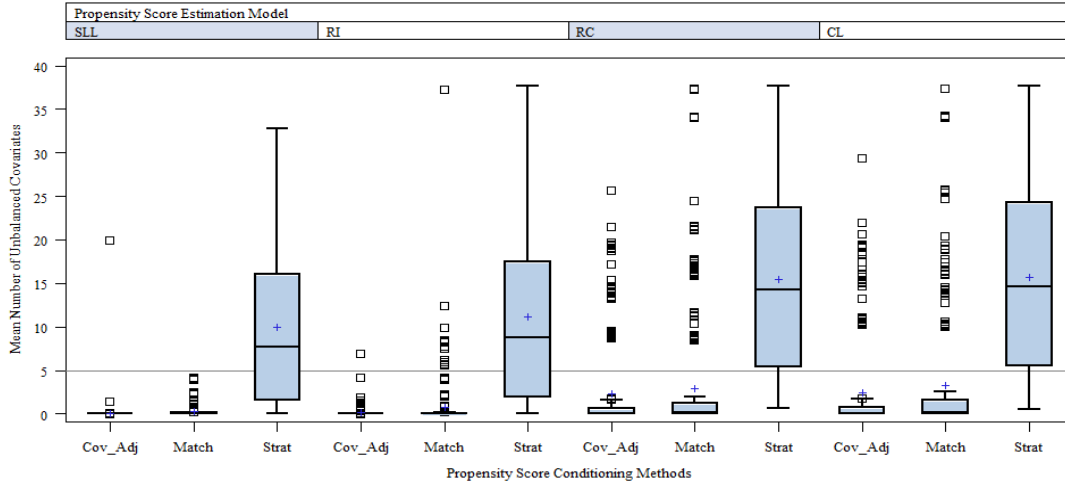


Figure 14. Mean number of unbalanced covariates by conditioning method across the four PS models

The variability in mean number of unbalanced covariates was further explored across the nine main effects and all first-order interactions. This model explained 93% of the variability in the mean number of unbalanced covariates. Factors associated with the variability in mean number of unbalanced covariates included conditioning method ($\eta^2=.36$), level-1 sample size ($\eta^2=.21$), the interaction between the conditioning method and the level-1 sample size ($\eta^2=.10$), and level-2 sample size ($\eta^2=.08$). Figure 15 displays the distributions associated with the interaction effect. This figure illustrates the impact small samples had on the mean number of covariates balanced. The stark contrast in the covariance adjustment and matching distributions between the small sample size level (1-9) and the other two levels is remarkable. The difference in distributions for covariance adjustment and matching may have been less drastic had all the samples for that level (1-9) converged.

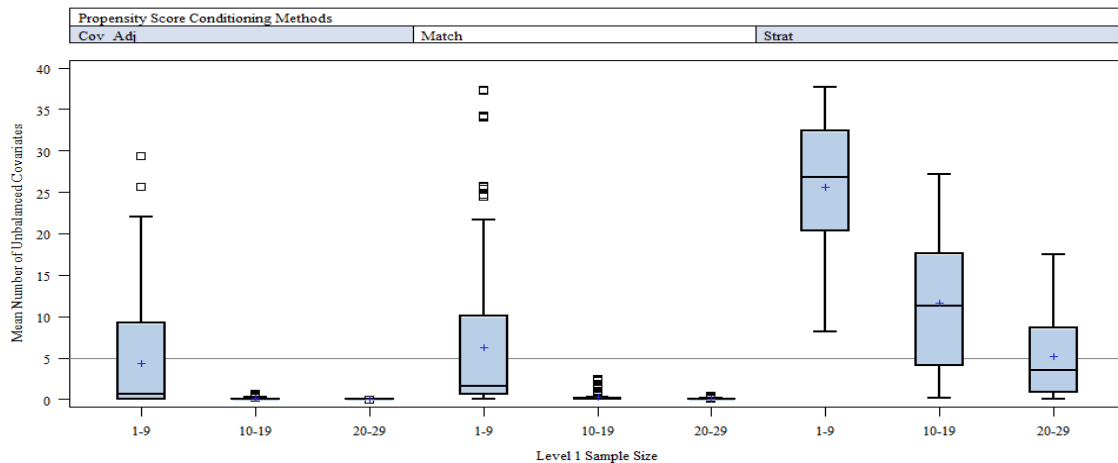


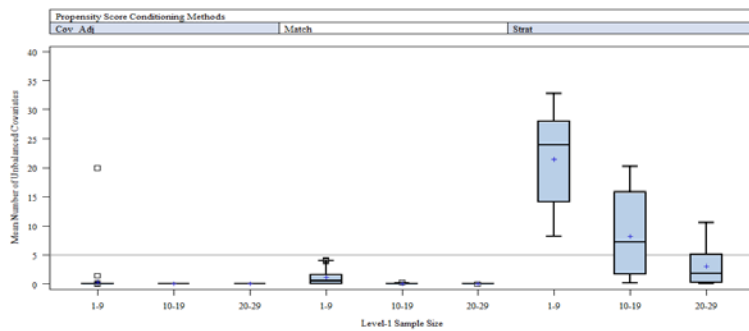
Figure 15. Distributions in mean number of unbalanced covariates by level-1 sample size across conditioning methods

The means and standard errors associated with this interaction effect are presented in Table 12. Stratification performed relatively poorly regardless of the level-1 sample size. However, when the level-1 sample size was 20-29, typical value of mean number of unbalanced covariates decreased and approached the threshold ($M=5.14$). Evidenced in Figure 15 and reiterated in Table 12 is the strikingly similar markedly impressive performance of covariance adjustment and matching on these data. On average, for moderate and large samples, both of these conditioning methods had almost no unbalanced covariates. Figure 16 displays the distributions of this interaction effect for each PS estimation model.

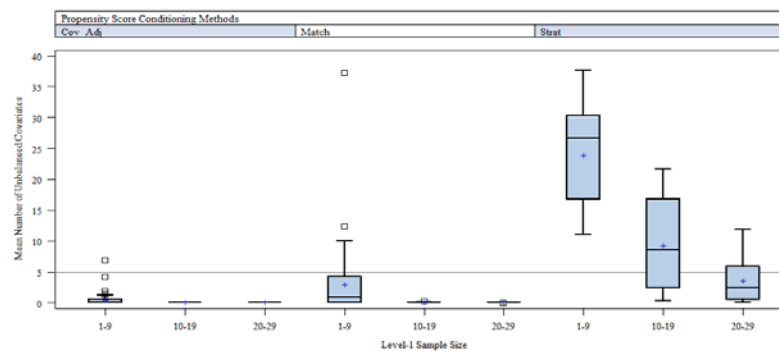
Table 12
Mean Number of Unbalanced Covariates by Level-1 Sample Size and Conditioning Method

Level 1 Sample Size	Conditioning Method					
	Covariance Adjustment		Matching		Stratification	
	Mean	SD	Mean	SD	Mean	SD
01-09	4.4	6.7	6.3	8.8	25.64	7.38
10-19	.12	.22	.30	.55	11.64	8.20
20-29	.004	.01	.05	.09	5.14	4.8

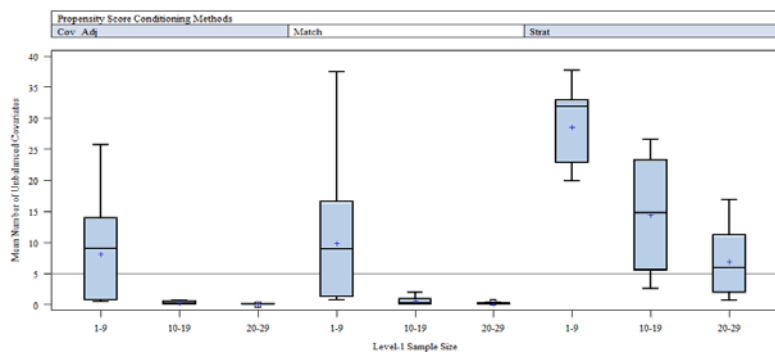
Single-Level Logistic



Random Intercepts



Random Coefficients



Cross-Level

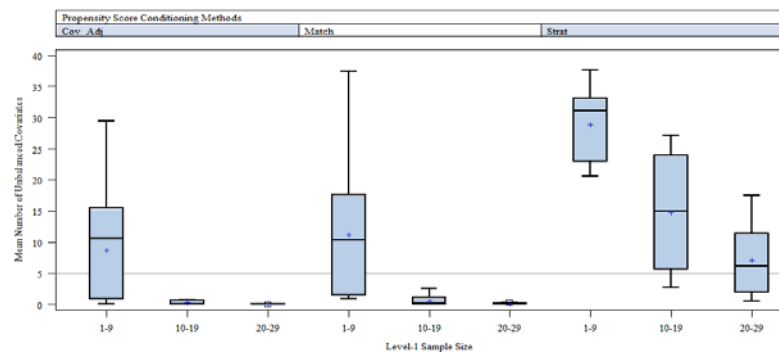


Figure 16. Distributions in mean number of unbalanced covariates by level-1 sample size across conditioning methods by propensity score estimation model

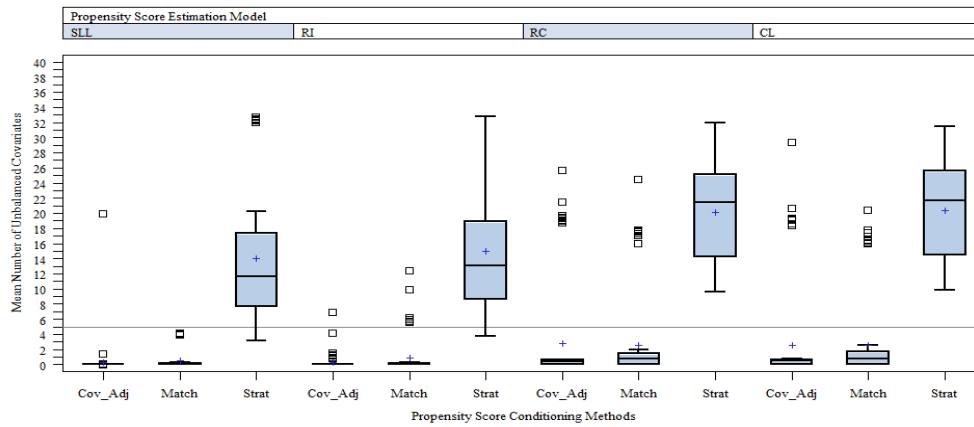
Descriptive statistics associated with the level-2 sample size main effect are presented in Table 13. These data indicate that as the number of clusters increases the typical mean number of unbalanced covariates decreases and the dispersion also decreases. For this main effect, when the number of clusters was 30, on average produced lower mean number of unbalanced covariates than when the number of clusters was 50. The large number of missing data due to the convergence issue only occurred when the number of clusters was 30; therefore it is likely that the missingness contributed to the smaller average and less likely that a smaller number of clusters would produce on average fewer unbalanced covariates. Figure 17 provides the distributions for this main effect across all PS methods (conditioning methods and estimation models).

Table 13
Descriptive Statistics for the Mean Number of Unbalanced Covariates by Level-2 Sample Size

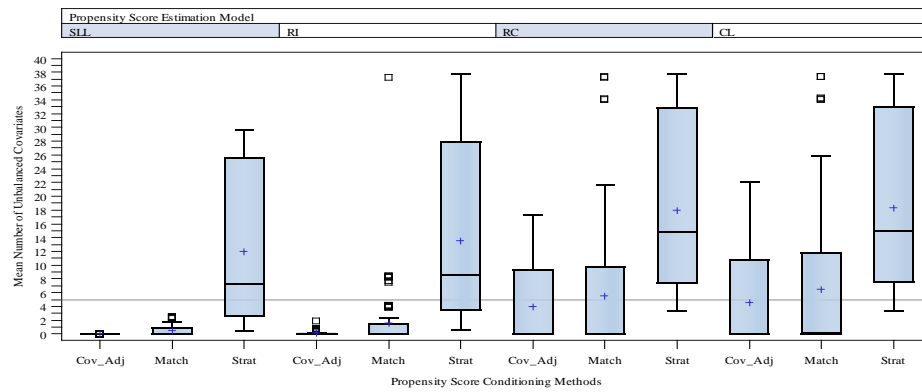
Level-2 Sample Size	Mean	SD	Min	Max
30	6.79	9.53	0	32.79
50	7.05	10.52	0	37.72
100	2.65	6.06	0	26.04

The third and final outcome measure used to evaluate balance in the study was the proportion of samples balanced. The previous variable counted the number of individual covariates yielding a balance score of greater than 0.25. A sample was considered unbalanced if more than 10% of the covariates were not balanced. Therefore, if the sample had 5 or more variables with a balance score of greater than 0.25 it was not considered balanced. Conversely, if the sample had 0, 1, 2, 3 or 4 variables that yielded a balance score of 0.25 or greater than it was considered balanced. Figure 18 displays the proportion of samples balanced for each conditioning method across the four PS estimation models. This figure highlights two distinct patterns. First, is the consistently large distributions and overall small proportion of samples balanced when data were stratified. Second, are the similarities in the distributions for the single level logistic model and the random intercepts model and the same similarities are seen with the random coefficients and cross-level models.

Number of Clusters=30



Number of Clusters=50



Number of Clusters=100

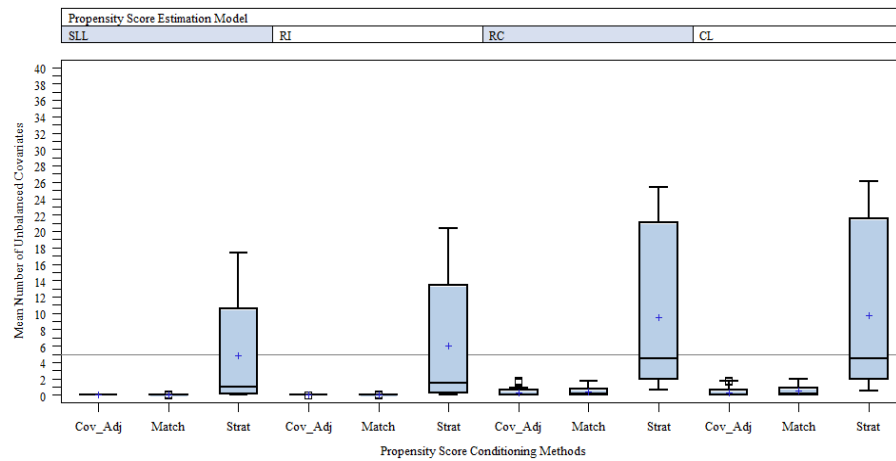


Figure 17. Distributions of mean number of unbalanced covariates by conditioning method across the four PS models for each level-2 sample size level

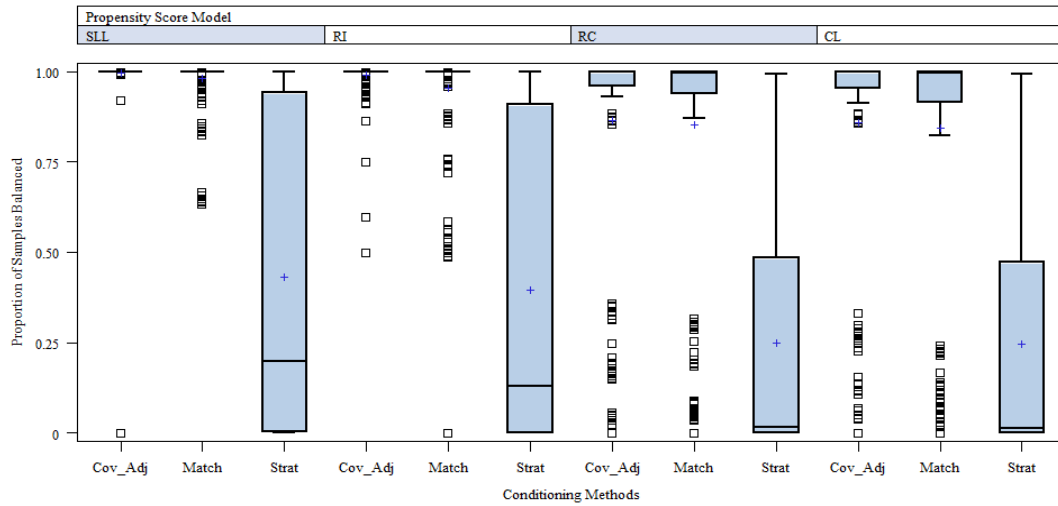


Figure 18. Proportion of samples balanced by conditioning method across PS models

The variability in the proportion of samples balanced was explored across the nine main effects and all first-order interactions. This model explained 85% of the variability in the proportion of samples balanced. The variability was further explored by including all second-order interactions. This model explained 97% of the variability in the proportion of samples balanced. Factors that were associated with this variability included conditioning method ($\eta^2=.46$), level-1 sample size ($\eta^2=.14$), level-2 sample size ($\eta^2=.08$), and the interaction across all the aforementioned factors ($\eta^2=.07$). Figure 19 displays this second-order interaction effect associated with the variability in the proportion of samples balanced. This figure depicts the consistent patterns with matching and covariance adjustment sharply contrasted with the inconsistency with stratification. There is drastic improvement with regard to the proportion of samples balanced when stratifying when level-1 sample size is 20-29 and the level-2 sample size is 100. Both matching and stratification produce large proportions of balanced samples when the level-2 sample size is 100, regardless of the level-1 sample size.

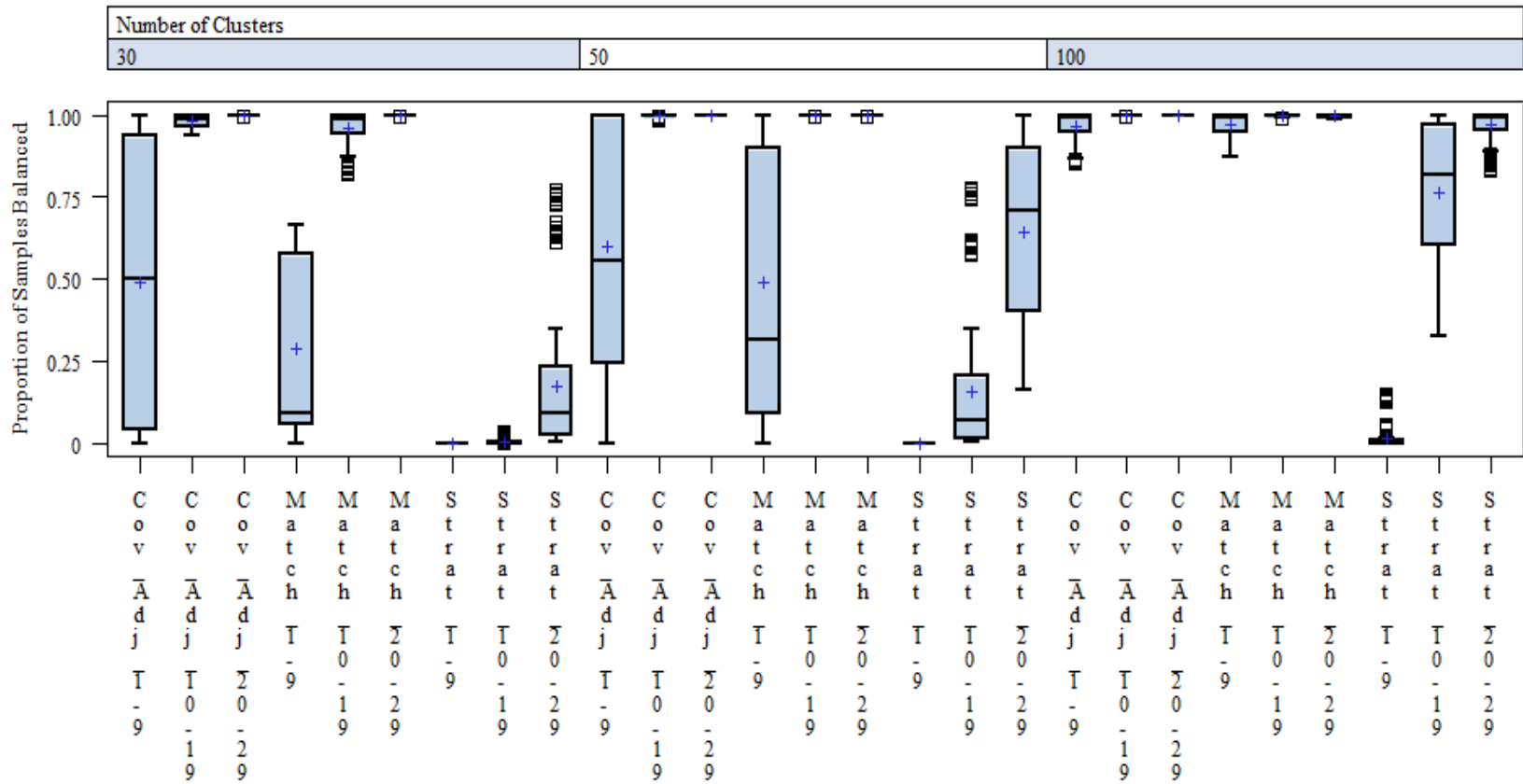


Figure 19. Proportion of samples balanced by conditioning method and level-1 sample size interaction across the level-2 sample size

Descriptive statistics for the second-order interaction associated with the variability in the proportion of samples balanced are presented in Table 14. As seen in this table, the data highlight the extreme performance of stratification, notably how when level 1 sample size was 1-9 and the cluster size was 50, no samples were balanced when stratifying. This is rather surprising as the convergence issue only occurred when level-1 sample size was 1-9 and the cluster size was 30; thus there were no missing data or a lack of samples for this cell. Covariance adjustment and matching performed remarkably well for seven out of the nine sample size interaction levels. Additionally all samples were balanced when using covariance adjustment when the level 1 sample size was 20-29 and the cluster size was 50 and 100.

Table 14
Descriptive Statistics for the Proportions of Samples Balanced by Level-1 Sample Size, Level-2 Sample Size and Conditioning Method

Level 2 Sample Size	Level 1 Sample Size	Conditioning Method					
		Covariance Adjustment		Matching		Stratification	
		Mean	SD	Mean	SD	Mean	SD
30	01-09	.49	.45	.29	.27	0	0
	10-19	.98	.02	.96	.05	.002	.006
	20-29	.99	.0002	.99	.0001	.17	.22
50	01-09	.60	.40	.49	.40	0	0
	10-19	.99	.003	.99	.0005	.15	.22
	20-29	1.0	0	.99	.003	.64	.27
100	01-09	.97	.04	.97	.03	.01	.03
	10-19	.99	.0002	.99	.001	.76	.21
	20-29	1.0	0	.99	.003	.97	.04

Treatment Effects

Four different variables were used to examine the effectiveness of the treatment effects estimates in the outcome model: statistical bias in the point estimates, the root mean squared error (RMSE), the 95% confidence interval coverage, and the 95% confidence interval width.

Bias in this study was calculated as the differences between the estimated treatment effect and the corresponding population parameter ($\delta=.2$ or $\delta=.5$) for each sample. The bias estimates were aggregated across replications and the distributions of the mean bias in the point estimates

are presented by conditioning method across the four PS models in Figure 20. Overall, the estimates across all conditioning methods and PS models were not very biased. The distributions got larger as the PS models became more complex consistently for all conditioning methods. For the single level logistic model, the typical values and dispersion of data were very small. Extremely biased values were seen in some conditions with the random coefficients and cross-level models for all conditioning methods.

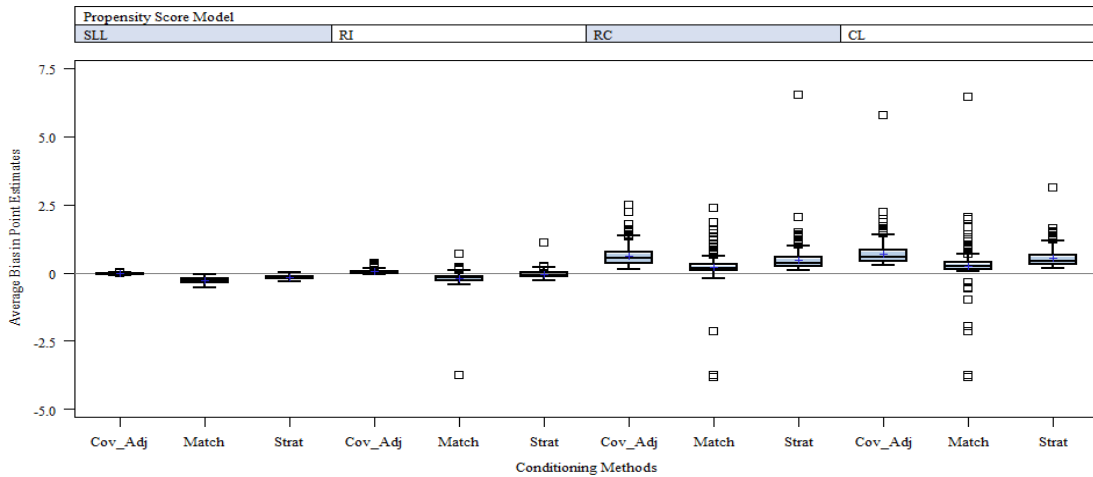


Figure 20. Distributions of estimated bias in point estimates by conditioning methods across PS models

The variability in bias was examined across all main effects and first-order interactions. This model only explained 67% of the variability, therefore all second-order interactions were included. The model with all main effects, first and second-order interactions explained 77% of the variability. Finally, a third model which included all main effects, first, second, and third-order interactions was considered. This final model explained a total of 87% of the variability in estimated bias. Neither the second nor the third model identified any additional factors related substantially to the variability in bias. Factors associated with the variability in bias included PS estimation model ($\eta^2=.36$), conditioning method ($\eta^2=.08$), and level-1 sample size ($\eta^2=.06$). Descriptive statistics for the three factors associated with the variability in bias are presented in Table 15.

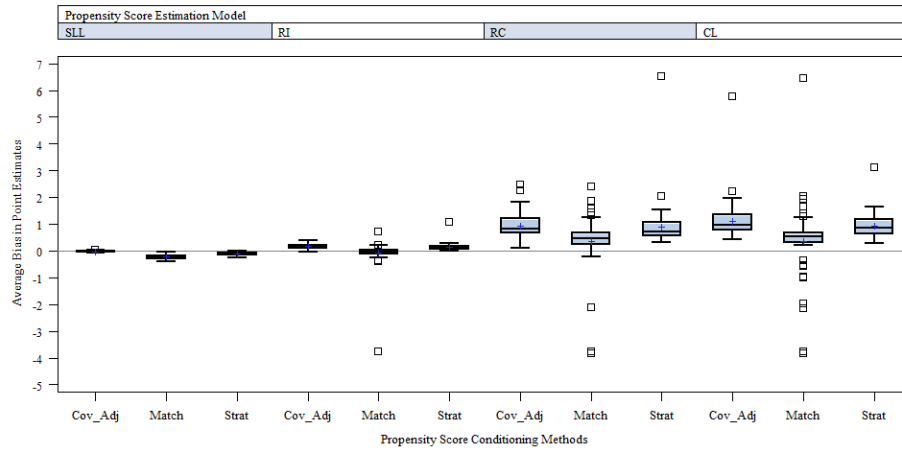
Table 15
Descriptive Statistics by Design Factors Associated with Bias

Design Factor	Mean	SD	Min	Max
PS Estimation Model				
SLL	-0.16	.12	-0.53	0.05
RI	-0.06	.20	-3.72	1.12
RC	0.43	.50	-3.79	6.55
CL	0.49	.55	-3.79	6.49
Conditioning Method				
Covariance adjustment	0.33	.44	-0.09	5.81
Matching	-0.001	.50	-3.79	6.49
Stratification	0.20	.40	-0.33	6.56
Level-1 Sample Size				
1-9	0.37	.76	-3.79	6.55
10-19	0.13	.31	-0.51	0.94
20-29	0.07	.29	-0.53	0.85

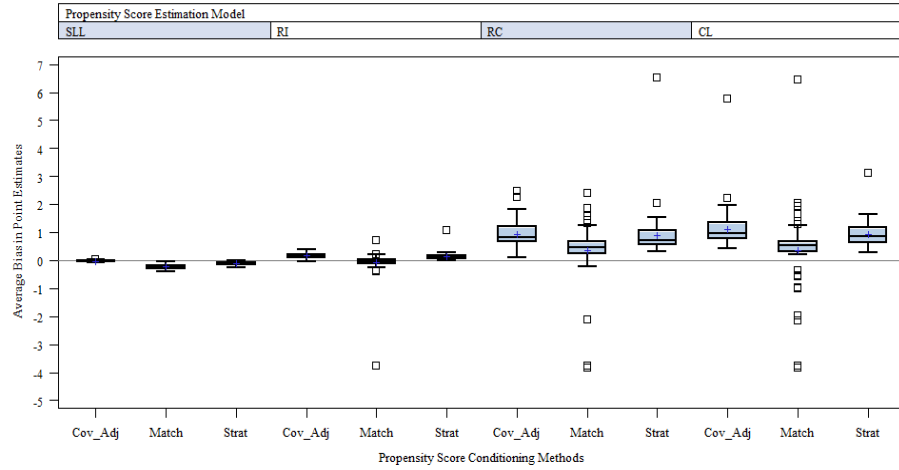
As the PS model became more complex the dispersion of the mean bias estimates increased. On average, the single level logistic and random intercepts model slightly underestimated the treatment effect while the random coefficients and cross-level models overestimated the treatment effect. With regard to the different conditioning methods, on average the mean bias estimate was very negligible for the matched samples. However, the range of bias estimates was the largest for this method. On average both stratification and covariance adjustment overestimated the treatment effect. The overall mean bias estimate decreased as the level-1 sample size increased. All three levels tended to on average overestimate the treatment effect. The range of bias estimates was largest when the level-1 sample size was 1-9. The range of bias estimates decreased tremendously for the remaining two level-1 sample size factors. Distributions for the level-1 sample size main effect across all PS methods (conditioning methods and estimation models) are provided in Figure 21.

The second outcome variable used to evaluate the effectiveness of the treatment effects estimates in the outcome model was the root mean squared error (RMSE). The RMSE was calculated by taking the square root of the sum of the differences between the estimated treatment effect and the corresponding population parameter divided by the total number of

Level -1 Sample Size =1-9



Level -1 Sample Size =10-19



Level -1 Sample Size =20-29

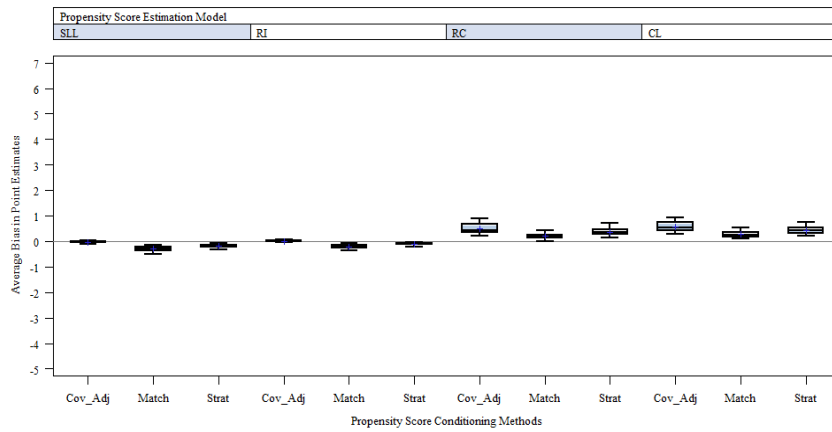


Figure 21. Distributions of the average bias in point estimates by conditioning method across the four PS models for each level-1 sample size level

samples converged. The distributions of RMSE estimates by conditioning methods across the four PS estimation models are presented in Figure 22.

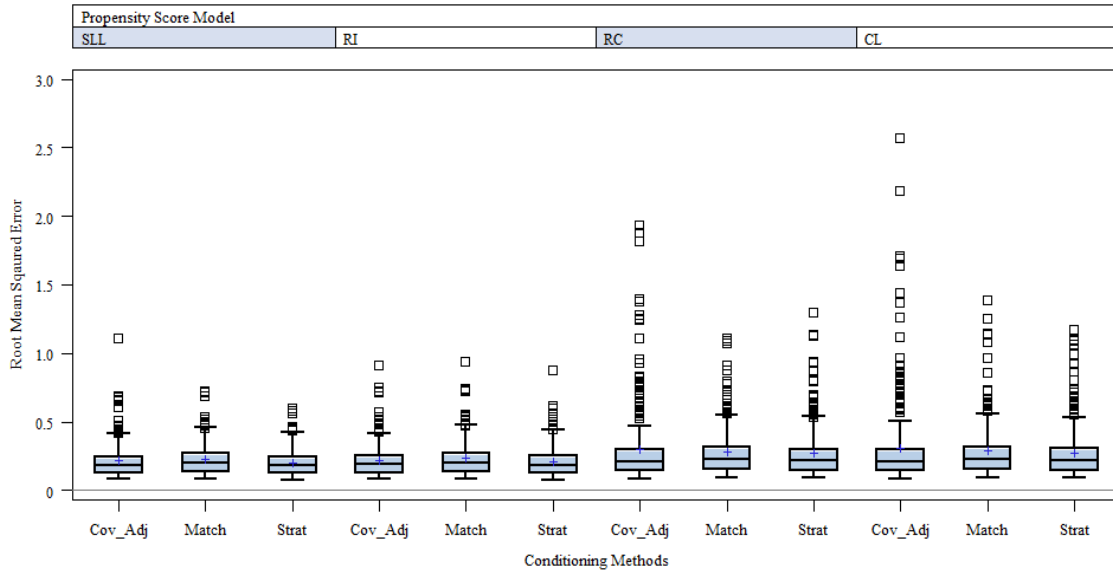


Figure 22. Distributions of the RMSE for each conditioning method across the PS estimation models

The distributions in RMSE values were quite consistent across conditioning methods for each PS estimation model. For the random coefficients and cross-level models relatively larger RMSE estimates are present, especially when using covariance adjustment. To explore the variability in RMSE estimates across the design factors two different models were considered. First the RMSE estimates were examined across all nine main effects and first-order interactions. This model explained 88% of the variability in RMSE values. The second model included all nine main effects, first and second-order interactions. This model explained 95% of the variability; however this second model did not identify any additional effects substantially associated with the variability in RMSE estimates. Factors associated with the variability in RMSE estimates included; level-1 sample size ($\eta^2=.41$), level-2 sample size ($\eta^2=.20$) and the interaction between the level-1 sample size and level-2 sample size. Distributions related to this interaction effect are presented in Figure 23.

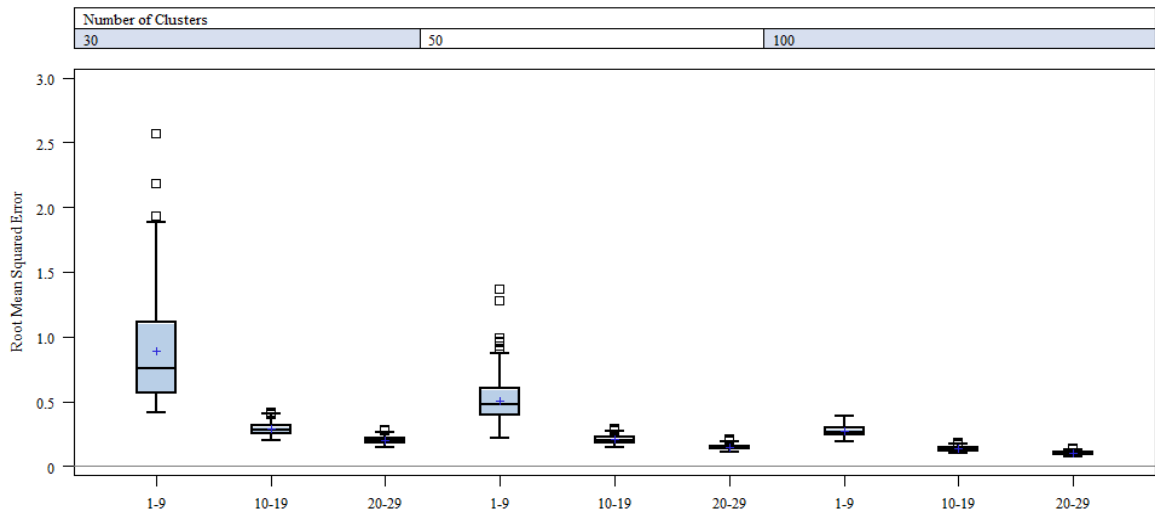


Figure 23. Root mean squared error for the level 1 sample size across the number of clusters

The RMSE was greater and more disperse when the level 1 sample size was small regardless of the number of clusters. The RMSE estimates were the largest when the level 1 sample size was 1-9 and the number of clusters was 30 or 50. The distributions of the RMSE estimates were quite small for all other sample size levels. Descriptive statistics for the RMSE estimates associated with this interaction effect are provided in Table 16. As the level-1 sample size increases the mean values of the RMSE estimates decrease consistently across level-2 sample size. Additionally, the same pattern is observed when the number of clusters increases across each level-1 sample size.

Table 16
Root Mean Squared Error by Level 1 Sample Size Across the Clusters

	Number of Clusters					
	30		50		100	
Level 1 Sample Size	M	SD	M	SD	M	SD
01-09	.89	.42	.51	.16	.27	.04
10-19	.29	.05	.21	.03	.14	.02
20-29	.21	.02	.15	.02	.10	.01

The third outcome variable used to evaluate the effectiveness of the treatment effects estimates in the outcome model was the 95% confidence interval coverage. The 95% confidence

interval coverage is the proportion of samples in which the population parameter specified ($\delta=.2$ or $\delta=.5$) fell within the 95% confidence interval. The distributions of coverage estimates across the conditioning methods for each PS estimation model are displayed in Figure 24. On average all the conditioning methods across all the PS estimation models under covered. As can be seen in the figure there are several conditioning methods where the upper quartile meets the threshold, and had adequate coverage rates. Across all PS estimation models, this was consistent with matching.

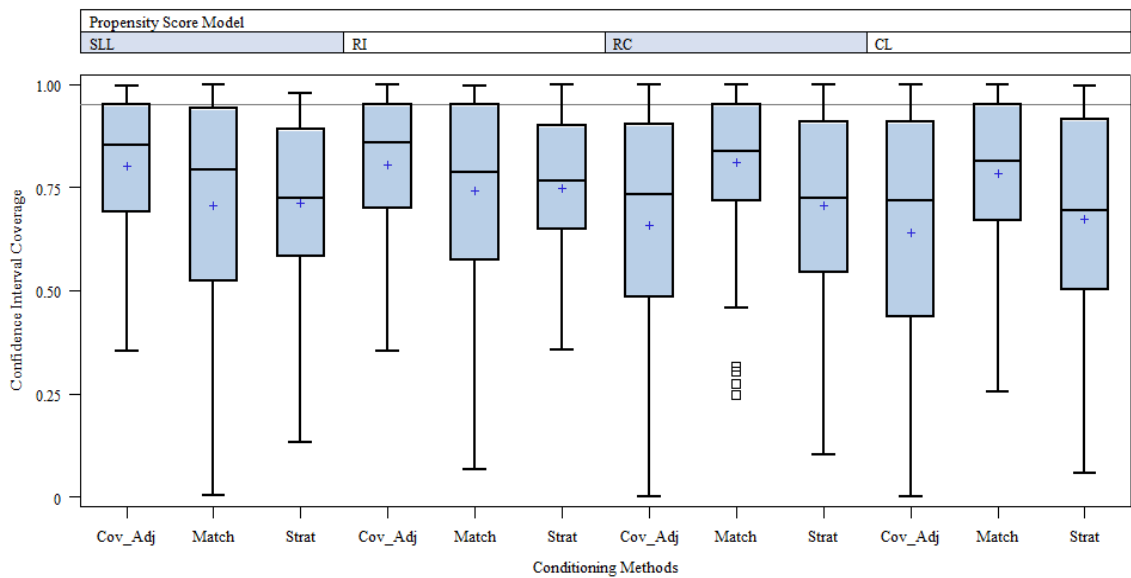


Figure 24. Distributions of 95% confidence interval coverage rates for each conditioning method across the four PS estimation models.

To explore the variability in the coverage rates across the design factors two different models were considered. First the coverage rates were examined across all nine main effects and first-order interactions. This model explained 67% of the variability in coverage rates. The second model included all nine main effects, first and second-order interactions. This model explained 85% of the variability. Factors associated with the variability in coverage estimates included level-2 sample size ($\eta^2=.22$), the level-1 sample size ($\eta^2=.17$), the interaction between the two aforementioned factors ($\eta^2=.11$), and the second -order interaction for conditioning method, estimation model, and level-2 sample size ($\eta^2=.06$).

The distributions illustrating the interaction between the level-1 sample size and the level-2 sample size factors are displayed in Figure 25. As the sample size increased, the distributions of coverage rates increased. This pattern was not consistent across all cluster levels. For example, when the level-1 sample size was 1-9 and the cluster size was 30, the median value of coverage rates showed adequate coverage; however, the convergence issues experienced likely contribute to the smaller distribution for this particular interaction level.

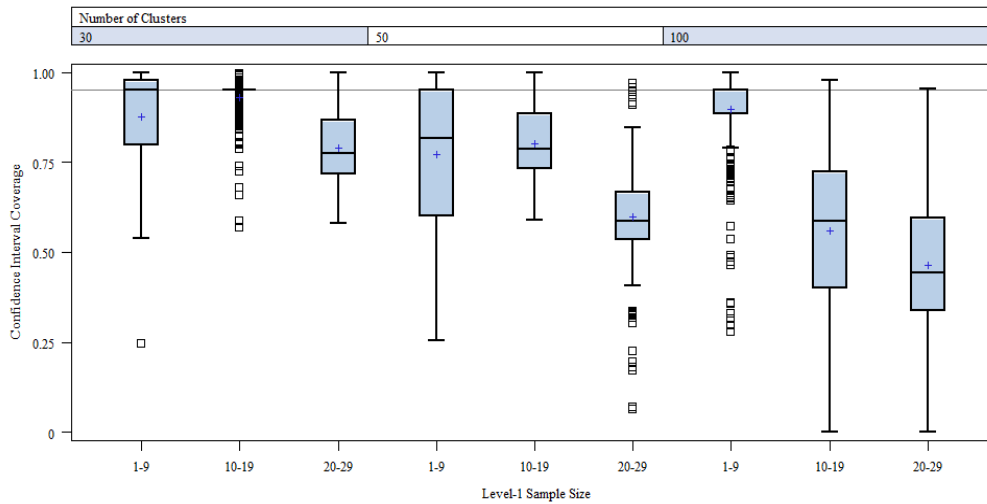


Figure 25. Distributions of 95% confidence interval coverage rates by level-1 sample size across the number of clusters.

The mean interval coverage estimates by the level-1 and level-2 sample size interaction effect are presented in Table 17. The data indicate the impact of the number of clusters is greater as the level-1 sample size increases. Across all three cluster values, the average coverage rate decreased substantially when increasing the level-1 sample size from 10-19 to 20-29.

The distributions illustrating the second-order interaction effect of the conditioning methods, PS estimation models, and level-2 sample size are presented in Figure 26. Evident in this figure is the increasing dispersion of coverage estimates as the number of clusters increased as well as the variability in coverage estimates for each conditioning method and PS estimation model. For example, with 100 clusters, the cross-level and random coefficients model both had wide distributions when using covariance adjustment with 100 clusters. The single level and

random intercepts model performed similarly when using matching with 100 clusters. In contrast, the distributions of the single level and random intercepts model were less dispersed when using covariance adjustment with 100 clusters. Additionally, the more complex models (random coefficients and cross-level) had smaller distributions of coverage estimates with matching for 100 clusters. The distributions of coverage estimates were fairly consistent across the estimation models and conditioning methods with 30 clusters.

Table 17
Confidence Interval Coverage by Level 1 Sample Size Across the Clusters

	Number of Clusters					
	30		50		100	
Level 1 Sample Size	M	SD	M	SD	M	SD
01-09	.88	.14	.77	.19	.90	.11
10-19	.93	.05	.80	.10	.56	.25
20-29	.80	.10	.60	.12	.46	.22

The mean interval coverage estimates by conditioning method, estimation model, and level-2 sample size are presented in Table 18. These data indicate that the impact of level-2 sample size is dependent upon the estimation model and conditioning method selected. For example, with covariance adjustment, mean coverage estimate increased when the clusters increased from 50 to 100 when using the single level or random intercepts models. However, the mean coverage estimates decreased from .68 to .46 and from .67 to .44 for the random coefficients and cross-level models respectively.

The fourth and final variable used to evaluate the effectiveness of the treatment effects estimates in the outcome model was the confidence interval width. The confidence interval width was calculated as the difference between the 95% confidence interval upper and lower limits. The distributions of confidence interval widths across the conditioning methods for each PS estimation model are displayed in Figure 27. The distributions suggest that the differences in

Table 18
Confidence Interval Coverage Estimates by Conditioning Method, Estimation Model and Level-2 Sample Size

PS Estimation Model	Conditioning Methods	Level-2 Sample Size					
		30		50		100	
		Mean	SD	Mean	SD	Mean	SD
SLL	Covariance	.85	.11	.77	.17	.79	.18
	Matching	.87	.10	.76	.18	.52	.30
	Stratification	.81	.12	.73	.16	.62	.22
RI	Covariance	.86	.11	.77	.16	.80	.18
	Matching	.87	.10	.76	.15	.62	.26
	Stratification	.82	.12	.72	.13	.72	.20
RC	Covariance	.88	.09	.68	.17	.46	.32
	Matching	.90	.11	.76	.14	.79	.17
	Stratification	.87	.10	.68	.15	.60	.24
CL	Covariance	.87	.11	.67	.19	.44	.33
	Matching	.90	.07	.73	.17	.75	.18
	Stratification	.87	.10	.66	.17	.54	.46

Note. Covariance refers to the covariance adjustment conditioning method.

average widths of the intervals across all conditioning methods for each model are very small.

Additionally, the variability in estimated confidence interval widths for each method is also relatively small, increasing with covariance adjustment for complex estimation models.

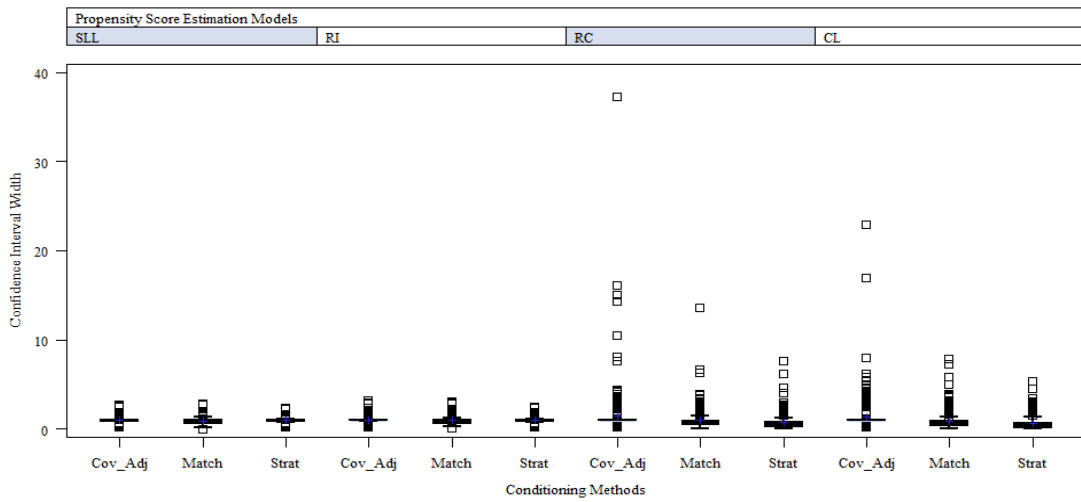


Figure 27. Distributions of confidence interval width by conditioning methods across PS estimation models.

To explore the variability in confidence interval widths three different models were considered. First, the variability in the confidence interval widths was examined across all main effects and first-order interactions. This model explained 48% of the variability in confidence interval widths. The second model included all main effects, first and second-order interactions. This model explained 67% of the variability. Finally, a third model was considered and included all main effects, first, second, and third-order interactions. Factors associated with the variability in confidence interval widths included: level-1 sample size ($\eta^2=.15$), the interaction between the level-1 and level-2 sample sizes ($\eta^2=.11$), the level-2 sample size ($\eta^2=.08$), and the second-order interaction of level-1 sample size, level-2 sample size and estimation model ($\eta^2=.07$). The distributions of confidence interval widths for this second-order interaction are displayed in Figure 28.

The distributions for the second-order interaction displayed in Figure 28 reveal that the differences in the average widths across the different levels was extremely small and the variability in the widths for each level was also small. The distributions of the cross-level and random coefficients models when level-1 sample size was 1-9 was relatively larger with 30 and 50 clusters. When the number of clusters was 100, there were almost no differences across the estimation method and level-1 sample sizes. Table 19 provides the means and deviations for each of the levels. The data revealed similar patterns for each PS estimation model. For example, the random coefficients and cross-level models had greater dispersion than the single level and random intercepts models.

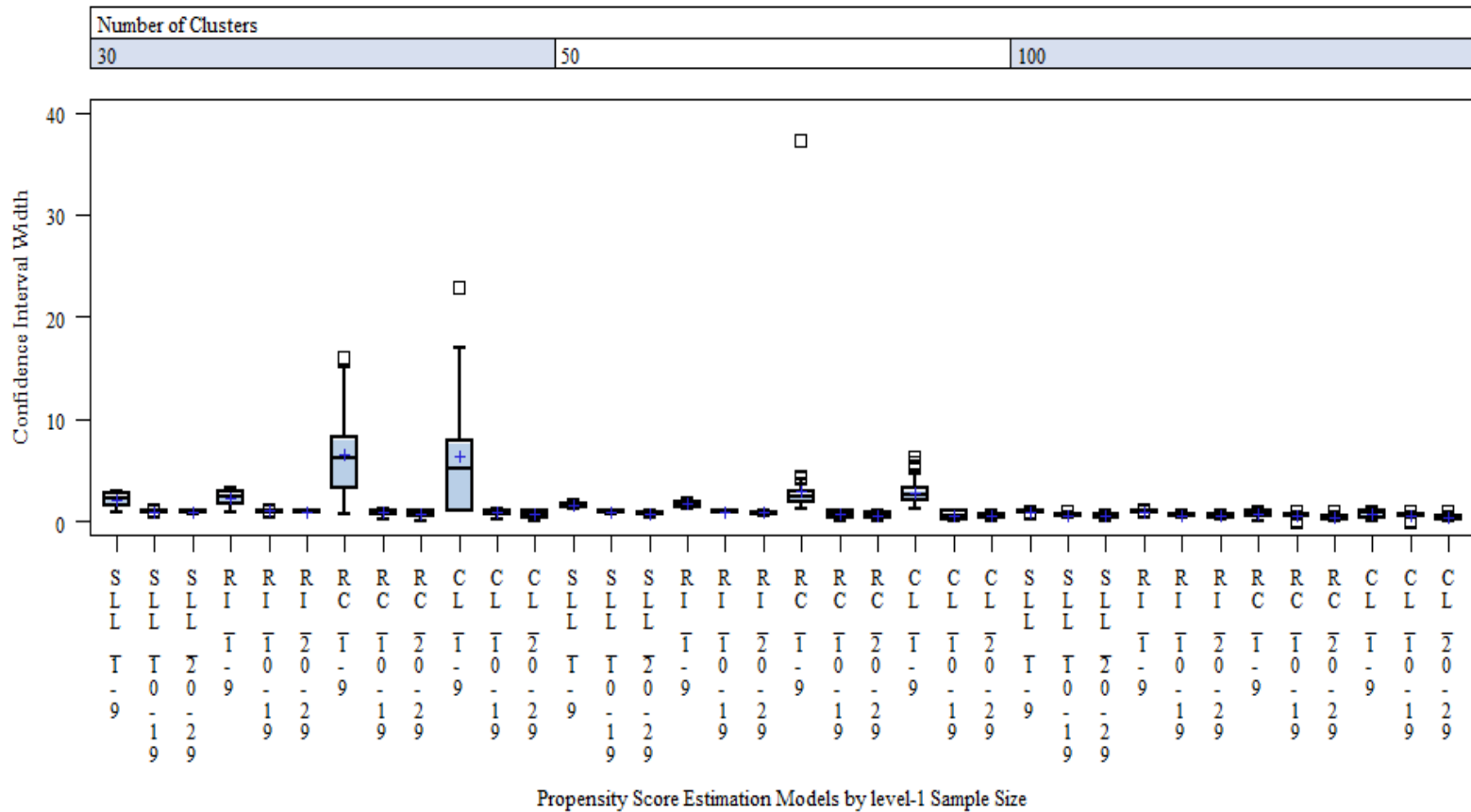


Figure 28. Distributions of the confidence interval width by propensity score estimation model and level-1 sample size interaction across the level-2 sample size

Table 19
Confidence Interval Width Averages and Distributions By Sample Size Interaction Effect Across Propensity Score Models

Number of Clusters	Level 1 Sample Size	SLL		RI		RC		CL	
		M	SD	M	SD	M	SD	M	SD
30	01-09	2.12	.67	2.27	.77	6.67	4.87	6.46	6.12
	10-19	.98	.05	.99	.04	.86	.24	.83	.27
	20-29	.93	.07	.94	.05	.74	.26	.68	.29
50	01-09	1.54	.22	1.7	.26	2.88	3.62	2.81	.99
	10-19	.93	.07	.96	.03	.65	.31	.59	.33
	20-29	.81	.16	.84	.14	.60	.31	.55	.33
100	01-09	.97	.04	.97	.05	.77	.33	.74	.37
	10-19	.59	.13	.62	.13	.54	.19	.54	.20
	20-29	.54	.25	.57	.23	.40	.27	.40	.27

Answers to Research Questions

The presentation above described the analysis procedures used to answer the first four research questions in this study. Presented below is a summary of how each of the analytical procedures provide answers to each of the first four research questions followed by a separate discussion regarding the final research question.

Research Question 1: To what extent do balance estimates vary across PS methods (PS estimation models and PS conditioning strategies)?

Recall, the three balance variables: the average absolute standardized mean difference for all 40 covariates, the average number of unbalanced variables, and the proportion of samples balanced. With regard to the first outcome measure, the average absolute standardized mean difference for all 40 covariates, the single level logistic model and random intercepts model performed better than the random coefficients and cross-level models. The performance of covariance adjustment and matching were similar. Covariance adjustment seemed to work better with the single-level logistic and random intercepts model, which matching outperformed covariance adjustment for the more complex PS estimation models. Stratification performed the

worst across all four PS estimation models with regard to average absolute standardized mean difference in covariates.

The second balance variable examined the number of unbalanced covariates in each sample. The PS estimation models all performed similarly as did covariance adjustment and matching across the four models. Stratification performed poorly across all four models and the average number of unbalanced covariates increased with model complexity.

The third variable examined the proportion of samples balanced. The single level logistic and random intercepts model performed better than the random coefficients and random intercepts model. Covariance adjustment and matching performed better than stratification. The performance of stratification was poor across all PS estimation models and progressively got worse as the PS estimation model became more complex.

Overall, the results revealed, the simpler PS estimation models may outperform the more complex PS estimation models marginally in terms of creating balanced groups. Additionally, the results indicate stratification does not do a good job creating balanced groups.

Research Question 2: To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the balance achieved by the PS methods (PS estimation models and PS conditioning strategies)?

Sample size had an impact on all three estimates of balance. In general, larger samples outperformed the smaller ones. Smaller average absolute balance score estimates were seen with the larger level-1 sample size, as well as with a greater number of clusters. The small level-1 sample size (1-9) had a large impact on the mean number of unbalanced covariates, especially with the more complex estimation models. Simply adding more level-1 cases decreased the mean number of unbalanced covariates substantially. The same was seen with the level-2 sample size. With 100 clusters, the difference in the mean number of unbalanced covariates was nearly identical across PS estimation models.

Finally, as the number of units within a cluster increased and the number of clusters increased, the proportion of balanced samples increased substantially. When the level-1 sample size was 20-29 and the number of clusters was 100, the majority of samples were balanced when using stratification. It was under this condition only that stratification performed comparably to covariance adjustment and matching.

Research Question 3: To what extent do treatment effect estimates vary across PS methods (PS estimation models and PS conditioning strategies)?

Recall, four different variables were used in this study to describe the treatment effect estimates, the bias in the point estimates, RMSE, the 95% confidence interval coverage, and the 95% confidence interval width. With regard to the bias in point estimates, the PS estimation models all performed similarly. The bias estimates tended to increase as the model became more complex. The conditioning methods also all performed very similarly. Matching performed the best, with the overall mean bias being the lowest, followed by stratification and finally covariance adjustment. The differences in conditioning methods were more pronounced as the model grew in complexity. For the single-level logistic and random intercepts models the differences in conditioning methods were almost negligible.

Neither propensity score estimation models nor conditioning methods had an impact on the RMSE estimates. However, more extreme values of RMSE were found across all conditioning methods for the random coefficients and cross-level models. However, with regard to the 95% confidence interval coverage estimates, both PS estimation models and PS conditioning strategies were related to their variability. Coverage rates dropped as the model became more complex. The decrease in coverage was greatest between the random coefficients and the cross-level model. Matching seemed to perform consistently with the random coefficients and cross-level models, where covariance adjustment performed better with the single-level logistic and random intercepts models. These were more pronounced as the level-2 sample size increased. Finally, there was very little variability in the estimates of confidence

interval width. Extreme values were present across all three conditioning methods with the random coefficients and cross-level models. The most extreme confidence interval width values were seen with covariance adjustment.

Research Question 4: To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the treatment effects estimated by the PS methods (PS estimation models and PS conditioning strategies)?

Sample size had an impact on all four treatment effect outcome measures. When the level-1 sample size was small the estimates tended to be more biased for the random coefficients and cross-level models across all conditioning methods. As the level-1 sample size increased to 20-29 the estimates of bias decreased and became less disperse.

The interaction between level-1 and level-2 sample size factors affected the estimates of RMSE and the 95% confidence interval coverage rates. Samples with 1-9 units within 30 and 50 clusters had larger and more disperse RMSE estimates. As the number of clusters increased to 100, the RMSE values decreased. The pattern was not as consistent with the confidence interval coverage estimates. In general, confidence interval coverage estimates were low. As the number of clusters increased the average rates dropped became more disperse. The variability in coverage rates increased as the sample size increased. The largest range of coverage rates occurred when there were 10-19 and 20-29 units within 100 clusters. The level-2 sample size impacted the coverage rates across the different PS estimation model and conditioning methods. The drop in the average coverage estimate was more pronounced when the number of clusters was 100. This was especially true for the cross-level model with all three conditioning methods. Finally, the level-1 sample size and level-2 sample size impacted the estimates of confidence interval width for the random coefficients and cross-level models only. With 1-9 units within 30 and 50 clusters the confidence interval widths were much wider than across any other PS estimation model.

Research Question 5: What is the direction and strength of the relationships between balance and both the accuracy and precision of the treatment effect estimates?

The literature on PS methods states that in order for one to consider the samples comparable to those obtained when randomly assigning units to treatment and control groups (i.e., a randomized experiment) the samples must be balanced. Previous research using Monte Carlo simulation methods for PSs using MLM has looked at the two independently. To answer this question the absolute value of the standardized mean difference between groups on the covariates, was used to represent the balance score. The results of this study were highly impacted by the sample size; therefore the relationships between balance and accuracy and precision of the treatment effect estimates were calculated controlling for both the level-1 and level-2 sample size.

Tables 20-24 provide the correlations between the balance estimates and the treatment effects, bias, RMSE, confidence interval coverage and confidence interval width respectively, by PS estimation model and PS conditioning method, partially out the effects of the level-1 and level-2 sample size.

Table 20
Correlations between balance score and absolute value of bias estimates for PS estimation models across PS conditioning methods

	Covariance Adjustment	Stratification	Matching
SLL	.05	-.19	-.25
RI	.61	.34	.77
RC	.50	.17	.67
CL	.69	.45	.71

The data in Table 20 reveal that there is a strong positive relationship between balance score and bias estimates for the MLM and covariance adjustment and matching. The strength of the relationships between balance score and bias for the MLMs was weaker for stratification than covariance adjustment and matching. The relationship between balance score and bias estimates for the single-level logistic model was negligible for covariance adjustment. There was a weak

inverse relationship between the balance score and the bias estimates for the single-level logistic model with the stratification and matching techniques.

Table 21
Correlations between balance score and RMSE estimates for PS estimation models across PS conditioning methods

	Covariance Adjustment	Stratification	Matching
SLL	.71	.77	.78
RI	.83	.66	.44
RC	.89	.59	.55
CL	.88	.56	.59

The data in Table 21 reveal that there in general strong positive relationship between balance score and RMSE estimates while controlling for the level-1 and level-2 sample size. The magnitude of the strength of the relationship varied across conditioning methods and PS estimation models. With covariance adjustment the relationship between balance score and RMSE got stronger with model complexity and was similar for the random coefficients and cross-level models. With stratification, the exact opposite pattern occurred. The strength of the relationship weakened as the model grew in complexity. Finally for matching, the relationship between balance score and RMSE was strongest with the single level model and weakest with the random intercepts model.

The data in Table 22 reveal, when controlling for the level-1 and level-2 sample size, there is an overall inverse relationship between balance score and the confidence interval coverage estimates, with the exception of the single-level logistic model and covariance adjustment. The strength of the relationship increases with model complexity. The differences in the magnitude of the strength of the relationship between the random coefficients and cross-level models are negligible for each conditioning method. The strength of the relationship between balance and confidence interval coverage was the strongest with matching.

Table 22

Correlations between balance score and confidence interval coverage estimates for PS estimation models across PS conditioning methods

	Covariance Adjustment	Stratification	Matching
SLL	.13	-.01	-.23
RI	-.08	-.18	-.31
RC	-.29	-.43	-.56
CL	-.29	-.45	-.60

Table 23

Correlations between balance score and confidence interval width for PS estimation models across PS conditioning methods

	Covariance Adjustment	Stratification	Matching
SLL	.54	.47	.78
RI	.77	.50	.52
RC	.52	.58	.66
CL	.79	.77	.79

Finally, the data in Table 23 identify a strong positive relationship between balance and confidence interval width controlling for the level-1 and level-2 sample size. In general, the strength of the relationships increased with model complexity, with a few exceptions. First, with covariance adjustment, the strength of the relationship decreased between the random intercepts and random coefficients model, before increasing again with the cross-level model. Second, with matching, the relationship between balance score and confidence interval width was strong with the single-level logistic model then decreased with the random intercepts model before increasing with the random coefficients and the cross-level models.

Chapter Summary

This chapter provided a description of the results of the study and answered the research questions. The PS estimation models, PS conditioning methods, the level-1 sample size and the level-2 sample size were factors that impacted the balance and treatment effect estimates. The following chapter provides an overview of the study, a discussion of these findings, the limitations of this study, as well as implications for future research.

CHAPTER FIVE: DISCUSSION

This chapter outlines a summary of the study and results, along with a discussion of the findings, limitations of the study, and implications for future research.

Summary of the Study

Purpose

The purpose of this study was to further examine the appropriateness of using PS methods to achieve balance between groups on observed covariates and to yield unbiased treatment effect estimates in multilevel studies. Specifically, this study examined the extent to which different PS approaches (PS estimation models and PS conditioning techniques) and sample characteristics (sample size, covariate relationship to treatment and outcome, and population effect size) achieved balance and reproduced the population treatment effect.

Research Questions

1. To what extent do balance estimates vary across PS methods (PS estimation models and PS conditioning strategies)?
2. To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the balance achieved by the PS methods (PS estimation models and PS conditioning strategies)?
3. To what extent do treatment effect estimates vary across PS methods (PS estimation models and PS conditioning strategies)?
4. To what extent do data factors (sample size, covariate relationship to treatment and outcome, and population effect size) affect the treatment effects estimated by the PS methods (PS estimation models and PS conditioning strategies)?

5. What is the direction and strength of the relationship between balance and both the accuracy and precision of treatment effect estimates?

Method

Monte Carlo simulation methods were used to examine the appropriateness of the methods. In addition to the two PS method factors (PS estimation models and PS conditioning methods) seven data factors were included. These factors were: (a) number of clusters (small [n=30], moderate [n=50], and large [n=100]); (b) within-cluster sample size (small [n =01-09], moderate [n =10-19], and large [n =20-29]); (c) relationship between level-1 covariates and treatment assignment (small [$\beta_{xz}=.10$], and moderate [$\beta_{xz}=.20$]); (d) relationship between level-1 covariates and outcome (small [$\beta_{xy}=.10$], and moderate [$\beta_{xy}=.20$]); (e) relationship between level-2 covariates and treatment assignment (small [$\gamma_{0s}=.20$], and moderate [$\gamma_{0s}=.40$]); (f) relationship between level-2 covariates and outcome (small [$\gamma_{s0}=.20$], and moderate [$\gamma_{s0}=.40$]); and (g) population effect size (δ = small [0.2] and moderate [0.5]). The values chosen for each of these factors were based on applied research to create realistic samples.

The data for this study were generated based on a cross level multilevel model (see Equation 19), using the PROC IML procedure in SAS (versions 9.2 and 9.3; SAS Institute, 2008). For each sample, four different PS models were estimated (see Appendix C) which yielded a PS for each unit of the sample. Samples were then trimmed to include only units falling within the region of common support between treatment and control groups. For each PS model, three different PS conditioning techniques were employed. Finally, balance estimates were calculated for each of the 12 PS estimation model/PS conditioning method combinations, and the outcome models were analyzed (see Equations 24, 26 and 28).

The results of the simulation were analyzed using PROC GLM in SAS 9.3 for both the balance and treatment effect estimates such that the dependent variables were balance score, number of unbalanced covariates, proportion of samples balanced, bias, RMSE, confidence

interval coverage, and confidence interval width and the independent variables were the two PS method factors and the seven data factors.

Discussion of the Study Results

Balance

The extent to which samples were balanced post PS estimation, trimming, and conditioning as a function of the PS estimation models, PS conditioning methods, and the seven sample characteristics was evaluated using three different outcome measures: the absolute value of the standardized mean difference between groups on the covariates, the mean number of unbalanced covariates, and the proportion of samples achieving balance.

The results indicate that regardless of sample condition, stratification fared noticeably worse than matching or covariance adjustment. These results are consistent with previous research comparing stratification and two different matching methods with nested data (Lingle, 2009). Lingle (2009) found that stratification did not perform as well as within cluster, or between-cluster matching when using significance testing to evaluate balance. However, the stratified samples retained a larger sample size, and results were interpreted cautiously as the non-significance could have been a result of balanced samples because of a reduction in sample size and power. The sub-optimal performance of stratification in terms of creating balanced groups as measured by estimating the standardized mean difference and not significance testing further corroborates previous findings. Additionally, this study found that balance was greater when using the single-level logistic and random intercepts model. These results are also consistent with previous research. Lingle (2009) found very few significant differences across all sample conditions when using a between-cluster matching approach for a single level logistic model, where balance was determined across the sample as a whole, rather than by each matched pair. In this study, balance was estimated for each matched pair, using a cross-classified random effects model. The consistency of the findings across the different methods of analysis (significance testing versus standardized mean differences, and whole group analysis versus matched pair)

suggest that stratification by quintiles may not be the best method to use in order to achieve balanced groups and between-cluster matching may be effective when using a single level logistic model to estimate PSs with nested data.

Previous Monte Carlo simulation studies examining PSs using MLM did not incorporate covariance adjustment as a conditioning method. Covariance adjustment differs in three aspects from other conditioning methods. First, conditioning and estimating the treatment effect occur in one step and thus does not require additional data loss (Steiner & Cook, 2013). Second, the covariate adjustment is the only conditioning method that uses a regression model relating the outcome to the treatment status and the PS. Finally, several regression assumptions apply when using covariance adjustment (Austin, 2011; Thoemmes & Kim, 2011). The overall impressive performance of covariance adjustment to produce balanced groups across all four PS estimation models is an important finding of this study.

Results also indicate that balance got appreciably worse as the PS model grew in complexity. That is, the single-level logistic model and the random intercepts model did appreciably better with regard to balance, than the random coefficients and cross-level models. These results were surprising considering the cross-level model was the closest to the generation model and previous research has found balance to be the greatest when models were specified accurately (Lingle, 2009; Thoemmes, 2009). Additionally, Kim and Seltzer (2007) used data from the EAOP and found including random effects of the slopes significantly improved the balance. These contradictory findings could be due to a number of reasons. For example, this study incorporated 30 level-1 covariates, and 10 level-2 covariates where covariates were correlated with each other and 10% of the covariates were dichotomous. Previous research has looked at a smaller number of continuous uncorrelated predictors, specifically three level-1 and one level-2 covariates (Lingle, 2009) and nine level-1 covariates and no cluster level predictor (Thoemmes, 2009).

Finally, this study found that balance was greatest when sample size increased. This finding is consistent with previous research (Lingle, 2009), where the significance rate dropped as the sample size increased, suggesting the differences between groups was minimal as the sample size increased. The impact of sample size on balance in this study was greater when using matching or stratification with a multi-level model. This is an important finding when trying to select a conditioning method for small samples that are nested.

Treatment Effects

The extent to which the samples were able to reproduce the population treatment effect as a function of the PS estimation models, conditioning methods, and seven sample characteristics was evaluated using four different outcome measures, bias, RMSE, 95% confidence interval coverage and 95% confidence interval width.

Overall, the results indicated that the estimates were largely unbiased with averages close and clustered around zero. The random intercepts model seemed to perform the best, in terms of being able to accurately reproduce unbiased treatment effect estimates with the single-level logistic model performing a close second. In contrast, the bias estimates were on average larger with the random coefficients and cross-level models. These results are interesting in comparison to previous research. Arpino and Mealli (2011) found that when omitting a potential unobserved cluster level predictor from very unbalanced samples, a fixed effects model worked well. Thoemmes (2009) found that relative to all the models, the single level model performed the worst in comparison to the multilevel models, and its performance depreciated as the ICC increased. The single-level models in previous research ignored the nesting structure completely as no level-2 covariates were included in the study. This current study incorporated the level-2 covariates into the single level model which may have accounted for the nested structure. It is uncertain whether the single-level logistic model would have produced larger biased estimates had the level-2 predictors been removed from the model. It is unclear as to why the random coefficients and cross-level models produced on average larger biased estimates than the random

intercepts model. Perhaps the amount of confounding in treatment effects was not strong enough. Thoemmes (2009) found that differences in models were larger under strong confounding conditions. Another reason for the random intercepts model performing better than the random coefficients or cross-level models would be if the random slope variance was not sufficiently large. This study attempted to avoid that problem by setting the random slope variance to .25 in order to induce a heavily nested data structure.

In terms of conditioning methods, on average, samples that used matching produced unbiased estimates. This is not surprising given the overwhelming support for and use of matching (Austin, 2008a, 2009, 2010; Imbens, 2004; Thoemmes & Kim, 2011). Stratification produced an average bias estimate of .20($SD=.40$), which was surprising. Literature often recommends stratification on the quintiles as a proven and effective way to remove 90% of the bias (Cochran, 1968; Rosenbaum and Rubin, 1984). However, previous research has not compared the performance of stratification in terms of reproducing the treatment effects to other conditioning methods using MLM. Few studies have compared the conditioning methods in a single level context and found the differences between the methods to be negligible.

Sample size was also an important contributing factor which impacted all four outcome measures associated with treatment effects. As the level-1 sample size increased, the bias estimates decreased. When the level-1 sample size was 20-29, estimates were, on average, unbiased. Additionally, the variability in RMSE across the different PS estimation models and conditioning methods was small; however, the variability in the estimates was associated with sample size. These findings are not surprising and are consistent with previous studies (Thoemmes, 2009) as well as with the broad methodological literature on sample size.

The extent to which the confidence intervals were accurate, as measured by the confidence interval coverage estimates, was dependent upon the PS estimation model, conditioning technique, and level-2 sample size. In general, the estimates tended to undercover. As the models became more complex and the samples increased in size the confidence interval

coverage estimates, on average, became less accurate. This was especially true for covariance adjustment and stratification. Matching performed relatively better than covariance adjustment and stratification with the complex models (random coefficients and random intercepts), while covariance adjustment performed relatively better than matching or stratification for the single-level logistic and random intercepts models with the larger samples. The mean proportion of confidence interval coverage did not hit 95% for any PS estimation model, conditioning method, or level-2 sample size condition. These results are not surprising as the smaller samples tended to be more biased, have a larger standard error, and wider confidence intervals. The overall poor coverage estimates are consistent with and slightly better than previous research, which results in almost 0% coverage for some models with larger samples (Thoemmes, 2009).

Finally, the width of the confidence interval was dependent upon the PS estimation model, level-1 sample size, and level-2 sample size. In general, the differences in the confidence interval widths across PS estimation models and conditioning methods were small. These small differences were associated with an interaction between PS estimation model, level-1 sample size, and level-2 sample size. When the level-1 sample size was 1-9 and the level-2 sample size was 30 or 50, the confidence intervals were less precise for all PS estimation models. The precision decreased noticeably with the random coefficients and the cross-level models. Conversely, for all other sample size conditions, the random coefficients and cross-level models were more precise than the single-level logistic and random intercepts models. Larger widths for smaller sample sizes are consistent with previous literature (Thoemmes, 2009) and the broad methodological literature on precision with inferential statistics.

The relationships between the estimated PSs were investigated through a series of correlations. These correlations were very high across all models, yet there were notable differences in the balance and treatment effect estimates across models. The most likely explanation for these seemingly paradoxical results is that the correlations were computed prior to the samples being trimmed and subsequently conditioned. The proportion of data trimmed, the

number of cases matched, and the proportion of strata retained all varied across the PS estimation models. The processes of trimming and conditioning the samples across the different estimation models likely led to the subsequent differences in balance and treatment effect estimates.

The Relationship Between Balance and the Accuracy and Precision of Treatment Effects

One of this study's contributions to the literature was the simultaneous investigation of the ability of PS estimation models and conditioning methods to create balanced groups and reproduce estimates of the treatment effect using MLM with clustered data. Previous literature examined the ability of different PS estimation models and conditioning methods either to create balanced groups (Lingle, 2009) or reproduce estimates of the treatment effect (Arpino & Mealli, 2011; Thoemmes, 2009; Thoemmes & West, 2011).

The purpose of the PS is to improve the quality of the estimates from non-randomized experiments by attempting to mimic the balance between groups that occurs through the randomization process (Rosenbaum & Rubin, 1984; Shadish & Steiner, 2010; Stuart, 2010). By creating balanced groups, in theory, the bias in estimates is thought to be removed, or at least, decreased. However, much of the methodological research does not investigate the nature of the groups as well as the quality of the estimates simultaneously. This study evaluated both the balance properties as well as the quality of the estimates; therefore it was prudent to investigate the relationship between the balance properties and the quality of the treatment effects. Estimates of bias and confidence interval coverage all indicate measures related to accuracy. The confidence interval width indicates the precision of the estimates, while RMSE incorporates both concepts of accuracy and precision. Since bias measures the extent to which the estimates over estimate or under estimate the population treatment effect, the absolute value of bias was used to calculate the relationship. The absolute average standardized mean difference in covariates was used to measure balance.

The results of this study were highly impacted by the sample size; therefore the relationships between balance and accuracy and precision of the treatment effect estimates were

calculated controlling for both the level-1 and level-2 sample size. Assuming that the relationship between balance and the accuracy and precision of the treatment effects is strong, a strong positive relationship between balance and bias, balance and RMSE, and balance and confidence interval width, and a strong inverse relationship between balance and coverage is expected.

Results indicate that balance was strongly related to the precision of the treatment effect estimates consistently across PS models and conditioning methods. Each PS estimation model and conditioning method had a strong positive relationship between balance and RMSE, and confidence interval width. The magnitudes of the relationships did vary across PS estimation models and conditioning methods; but all ranged from moderately strong to very strong.

With regard to accuracy, the strength and direction of the relationship varied across PS estimation model and conditioning method. The relationship between balance and bias for the single-level logistic model was interesting. With covariance adjustment, the relationship was positive albeit almost negligible. The relationship between balance and bias for the single-level logistic model for stratification and matching was surprisingly inverse however, not very strong. Additionally, the overall strength of the relationships between balance and the confidence interval coverage estimates were not as strong for the three MLM models as they were with balance. The direction of the relationship was contradictory to what was expected with the single-level logistic model with covariance adjustment. These results of this study, specifically, the analysis of the relationship between balance and the precision of treatment effect estimates confirm the theoretical framework of PSs.

Limitations of the Study

Based on the design of this study, there are generalizability limitations to consider with regard to this research study. This study incorporated Monte Carlo simulation methods to examine the performance of PS methods with MLM. Simulation methods allow for the control and manipulation of specific design and data factors to investigate the behavior of statistical methods (Guo & Fraser, 2010); while this is a benefit to simulation research, it also limits the

generalizability of the findings. Thus, the levels of the nine design factors (the two PS methods and the seven data factors) determine the scope to which the study's findings can be generalized.

First, the seven data factors play a large role in the generalizability. These levels were chosen to represent realistic sample characteristics found in education research. However, these levels are not exhaustive of all the possible values. Another limitation to consider relates to the PS models under investigation. Four different PS models (see Appendix C) were used to estimate the scores which made several assumptions. First, all models maximized the information available in the covariate set. The covariate set in this study was also very specific, with 30 (twenty-seven continuous and three dichotomous) level-1 predictors, and 10 (nine continuous and once dichotomous) level-2 predictors with correlations among the predictors within a level being fixed at 0.2. Continuous covariates were standardized with a mean of 0 and a variance of 1.0, while dichotomous predictors were generated from a binomial distribution where the population mean is approximately 0.5. The fixed effects for the intercepts were set to 1.0. The level-1 and level-2 errors were generated from a normal distribution with the level-1 error variance set at 1.0 and the level-2 error variance of .25 which produced conditional ICCs of .20. Finally, samples were generated using a variance components covariance structure where all covariance parameters in the random effects matrix were fixed to 0. This data generation process created a linear relationship between infallible predictors and the outcome. Additionally, only cross-level interactions were simulated in the study. Had the relationships between predictors and outcome been non-linear, had there been measurement error in the predictors, had there been interactions between two or more level-1 predictors, or any combination of these, the results obtained from the four PS estimation models used in this study may have been more biased and the confidence interval coverage rates may have decreased.

The performance of the PSs, especially the random coefficients and cross-level model may have been affected by the large number of covariates. Previous research found the single-level models performed poorly as compared to the multilevel models. However, the study

included nine level-1 covariates, no level-2 predictors, and two random slopes. The large number of covariates surely impacted the convergence problems that occurred with the small samples. The increase from 9 to 30 level-1 covariates, zero to ten level-2 covariates, and two to five random slopes may have impacted the performance of the complex models, especially the cross-level model, which was the model closest to the data generation model. This study did not include a prime facie ANCOVA model that mimicked the data generation process as a reference condition.

A third limitation to the study relates to the three different conditioning methods used. Each conditioning method includes multiple options and the results of this study are limited to the specific variations delimited (see Table 2). For example, when using matching, several different methods could be selected such as the type of matching algorithm, the caliper width, and with MLM, whether to match within or across clusters. In addition, stratification can be done with a varying number of strata. This study stratified the samples into quintiles; however including more strata may result in better performance, especially with regard to balance.

Fourth, the variables used to operationalize balance can also be considered a limitation of this study. With regard to PS analysis, a general guideline or cut off to determine the degree of balance between treatment and control groups has not been established. While the standardized mean difference score is widely used measure for continuous covariates, the method to convert differences between dichotomous variables to a continuous scale is still under investigation (Kromrey & Bell, 2012). Additionally, literature suggests that before estimating the treatment effects, balance should be checked and if adequate balance has not been achieved, then the PS estimation model should be modified, by including more or fewer covariates, considering non-linear or non-additive models. This study did not evaluate the balance samples prior to estimating the treatment effects.

Implications

Since the early 80's researchers have studied PS methods as well as multilevel models individually with respect to their ability to estimate effectiveness. Many social and behavioral studies employ non-randomized studies in hierarchical settings to estimate treatment effects. Both PS methods and multilevel modeling have advantages with regard to the social science research settings. The results of this study suggest that the degree to which using PSs with MLM is appropriate to create balanced groups and estimate the treatment effects is dependent upon the PS methods (estimation models and conditioning methods, as well as sample size. These findings led to implications for both methodologists and those who wish to conduct PS analysis with MLM.

Implications for Researchers Conducting PS Analysis with MLM

For researchers conducting PS analysis with MLM, the results of this study provide a few recommendations. First, it should be noted that the results of this study do not conclusively suggest that MLM PS estimation models work better than a single-level logistic regression model when data are clustered. In fact, the results of this study do not conclusively identify a superior PS estimation model or condition method or interaction between the two. What the results do suggest is that samples with fewer than 10 units nested within 30 or fewer clusters, trying to estimate PSs using a vector of a large number of covariates will likely not be a large enough sample to provide balanced samples, accurate and precise estimates, if the sample even converges.

With the PS MLM, the relationship between the precision of the estimates and the balance was strong and positive and therefore confirm the necessity of considering balance estimates when using PS methods. When samples are unbalanced then perhaps reconsider the estimation model, or the covariates used. This study maximized the information criteria for the covariate set for all PS estimation models, which in applied research is impossible to guarantee.

Finally, PS with MLM is still a novel methodological concept which requires a great deal of additional research. Conducting PS analysis using MLM when data are clustered in nonrandomized studies will definitely help the field, but findings from future studies using PS methods with clustered data should be interpreted and regarded with caution. This study did not identify consistently specific conditions where the PS estimation models and conditioning methods worked well in creating balanced groups and estimating the treatment effects accurately and precisely. At times, the single-level logistic model and random intercepts model gave the illusion of superiority with smaller samples (as evidenced by the higher percentage of confidence interval coverage estimates), but these averages are based on converged samples only. It is important to apply PS methods with clustered data to compare and evaluate the estimates from a methodological perspective; however, at this time it is not recommended to apply PS methods with clustered data in order to make high stakes decisions, or draw causal inferences.

Implications for Methodologists

For methodologists studying the use of PS methods with multilevel data, more research needs to be conducted on situations where the PS MLM model does not fully maximize the information of the covariate set. For example, this could include covariates with different magnitudes of the relationship between treatment being omitted from the PS estimation model. Additional research using different PS MLM models need to be investigated. This could include PS models that estimate the parameters of the covariance matrix or violate assumptions (e.g. nonnormality of the level-1 or level-2 errors, and heteroscedasticity of errors at all levels). Investigations of additional PS MLM models would allow for a better understanding of the applicability of the models to a variety of conditions.

Future research on different variations of conditioning methods should also be considered. For example, using different matching algorithms, calipers, both within and across clusters would also add to our understanding of the nature of PS methods with MLM. This study utilized a between cluster matching approach which thus made the outcome model a cross-

classified random effects model. Perhaps investigating the differences among the different variations of matching would also be worthwhile as this is a conditioning technique often employed in other disciplines (e.g. medicine).

Propensity score methods were originally created as a method to produce balanced groups, akin to those that occur naturally when employing the process of random assignment. However, much of the research on PSs within a single level context focuses on the ability of the methods to reproduce the treatment effect. Lingle's (2009) study focused specifically on the ability of a PS MLM to balance groups using three covariates. While that seminal research was necessary, it is not unheard of to use three covariates individually in the outcome model, and reducing three covariates to a single score may not be the most effective method. Balance is important, in order to make causal inferences similar to those being made when using random assignment. Stratification did not perform as well as expected with regard to creating balanced groups. As previously mentioned this could have been due to the aggregation of balance estimates across strata. Perhaps future research could investigate the individual strata to see if the scores within each stratum are comparable across all five strata. Further research in this area is required even with PS in the single-level context.

Finally, this study focused on the appropriateness of PSs to produce balanced groups and reproduce estimates of the treatment effect using data that were generated based on a two level random effects model. It would be interesting to investigate the appropriateness of different PS models and methods when data are further clustered in 3 or more levels, or even cross classified.

REFERENCES

- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.
- Asparouhov, T. (2004). Weighting for unequal probability of selection in multilevel modeling, Mplus Web Notes No. 8 available from <http://www.statmodel.com/>.
- Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics and data analysis*, 55, 1770-1780.
- Austin, P.C. (2007). The performance of different propensity score methods for estimating marginal odds ratio. *Statistics in Medicine*, 26, 3078-3094.
- Austin, P. C. (2008a). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2037-2049.
- Austin, P.C. (2008b). The performance of different propensity score methods for estimating relative risks. *Journal of Clinical Epidemiology*, 61, 537-545.
- 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. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology*, 172(9), 1092-1097.
- Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.

- Austin, P.C., Grootendorst, P., & Anderson, G.M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, 26, 734-753.
- Austin, P.C., Grootendorst, P., Normand, S.L., & Anderson, G.M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine*, 26, 754-768.
- Austin, P.C., & Mamdani, M.M. (2005). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, 25, 2084-2106.
- Bell, B.A., Ferron, J.M., Kromrey, J.D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *Proceedings of the Annual Joint Statistical Meeting*, 1122-1129.
- Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149-1156.
- Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1), 31-72.
- Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings, *Psychological Bulletin*, 54(4), 297-312.
- Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: RandMcNally.
- Chantala, K., Blanchette, D., & Suchindran, C.M. (2011). Software to compute sampling weights for multilevel analysis.
- Chatterji, M (2008). Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher*, 37(1), 23-26.

- Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19(3) 3127-3131.
- Cochran, W.G., (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, 128- 234-255.
- Cochran, W.G., (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2), 295-313.
- Cochran, W.G., & Rubin, D.B. (1973). Controlling bias in observational studies: A Review, 35, 417-446.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Lawrence Erlbaum Associates.
- Cook, T.D. (2006). Describing what is special about the role of experiments in contemporary educational research: Putting the "Gold Standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation*, 6, 1-7.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: RandMcNally.
- Cook, T. D., and Steiner, P.M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of covariate choice, unreliable measurement, and mode of data analysis. *Psychological Methods*, 15(1), 56-68.
- Dai, J., Li, Z. and Roche, D. (2006). Hierarchical logistic regression modeling with SAS GLIMMIX. University of California, Davis, CA, USA.
- Dedrick, R.F., Ferron, J.M., Hess, M.R., Hogarty, K.Y., Kromrey, J.D., Lang, T.R., Niles, J.D., & Lee, R.S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69-102.
- Dehejia, R.H. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics*, 125, 355-364.

- Dehejia, R.H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of American Statistical Association*, 94(448), 1053-1062.
- Dehejia, R.H., & Wahba, S. (2003). Propensity score matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1), 151-161.
- Donner, A. & Klar, N. (2000). *Design and analysis of cluster randomization of trials in health research*. John Wiley & Sons.
- Education and Sciences Reform Act (2002) Public Law NO. 107-279. 107th Congress, 107 *Congressional Record*, 147, 16 Stat.
- Fan, X., Felsovalyi, A., Sivo, S.A., & Keenan, S.C. (2002). *SAS® for Monte Carlo studies: A Guide for quantitative research*. Cary, NC: SAS Institute, Inc.
- Ferron, J.M., Hogarty, K.Y., Dedrick, R.F., Hess, M.R., Niles, J.D., Kromrey, J.D. (2008). *Reporting results from multilevel analyses*. In A.A. O'Connell; & D.B. McCoach (Eds.) *Multilevel Modeling of Educational Data* pp. 391-426.
- Gall, M.D., Gall, J.P., & Borg, W.R. (2007) *Educational Research: An Introduction* (8th Ed). Boston: Pearson.
- Games, P. (1990). Correlation and causation: A Logical snafu. *Journal of Experimental Education* 58(3), 239-246.
- Greenland, S. (2008). Invited Commentary: Variable selection versus shrinkage in control of multiple confounders. *American Journal of Epidemiology*, 59, 819-828.
- Gu, X.S., & Rosenbaum, P.R. (1993). Comparison of multivariate matching methods; Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4). 405-420.
- Guo, S., & Fraser, M.W. (2010). *Propensity Score Analysis: Statistical methods and applications*. London: Sage.

- Hahs-Vaugh, D.L., & Onwuegbuzie, A.J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education*, 75(1), 31-65.
- 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), 234-249.
- Heckman, J. & Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics*, 86, 30-57.
- Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46(3), 477-513.
- Hirano, K. & Imbens, G. (2002). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2, 259-278.
- Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161-1189.
- Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analyst*, 15(3), 199-236.
- Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945-970.
- Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

- Hong, G., & Raudenbush, S. W. (2003). "Causal inference for multi-level observational data with application to kindergarten retention study," 2003 Proceedings of the American Statistical Association, Social Statistics Section [CD-ROM, pp.1849-1856], Alexandria, VA: American Statistical Association.
- Hong, G., & Raudenbush, S.W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27(3), 205-224.
- Hong, G., & Raudenbush, S.W. (2006). Evaluating Kindergarten retention policy: A case study for causal inference for multilevel observational data. *American Statistical Association*, 101(475), 901-910.
- Hong, G., & Yu. B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, 44, 407-421.
- Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hughes, J.N., Chen, Q., Thoemmes, F., Kwok, O., (2010). An investigation of the relationship between retention in first grade and performance on high stakes tests in 3rd grade. *Educational Evaluation and Policy Analysis*, 32(2), 166-182.
- Imai, K., King, G., & Stuart, E.A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A* 171(2), 481-502.
- Imbens, G. W. (2003) Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, 93(2), 126-132.
- Imbens, G.W. (2004). Nonparametric Estimation of Average Treatment Effects: A Review. *Review of Economics and Statistics* 86, 4-29.

- Kelcey, B. (2011). The Role of Outcome Proxies in Propensity Score Variable Selection. *Multivariate Behavioral Research, 46*(3), 453-476.
- Kim, J., & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection processes vary across schools*. Center for the Study of Evaluation Technical Report 708.
- Kincaid, C. (2005). Guidelines for selecting the covariance structure in mixed model analysis. *Proceedings from the Annual Meeting of the SAS Users Group International, 198* (30), 1-8.
- Kromrey, J.D., & Bell, B.A. (2012). Effect size indexes for dichotomized outcomes under variance heterogeneity: An empirical investigation of accuracy and precision. In *JSM Proceedings, Social Statistics Section*. Alexandria, VA: American Statistical Association.
- Kurth, T., Walker, A.M., Glynn, R.J., Chan, K.A., Gaziano, J.M., Berger, K., & Robins, J.M. (2005). 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.
- LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*, 604-620.
- Lee, B.K., Lessler, J., & Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*, 337-346.
- Lingle, J.A., (2009). Evaluating the performance of propensity scores to address selection bias in a multilevel context: A Monte Carlo simulation study and application using a national data set. Unpublished Dissertation. AAT 3410727.
- Luellen, J.K., Shadish, W.R., Clark, M.H. (2005). Propensity Scores: An introduction and experimental test. *Evaluation Review 29*(6), 530-558.

- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via propensity score in estimation of causal treatment effects: A comparative study. *Statistical Medicine*, 23, 2937-2960.
- Maxwell, S.E. (2010). Introduction to the special section on Campbell's and Rubin's conceptualizations of causality. *Psychological Methods*, 15(1), 1-2.
- 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.
- McCoach, D.B. (2010). Hierarchical Linear Modeling. In G.R. Hancock & R.O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* pp 123-140.
- Michalopoulos, C., Bloom, H.S., & Hill, C.J. (2004). Can propensity score methods match the findings from a random assignment evaluation of mandatory welfare to work programs? *The Review of Econometrics and Statistics*, 86(1), 156-179.
- Ming, K., & Rosenbaum, P.R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computations and Graphical Statistics*, 10(3), 455-463.
- Morgan, S.L. & Todd, J.J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* 38, 231-281.
- Murphy, S.C., Law, S., Whooley, B.P, Alexandrou, A., Chu, K.M., & Wong, J. (2003). Atrial fibrillation after esophagectomy is a marker for postoperative morbidity and mortality. *Journal of Thoracic and Cardiovascular Surgery*, 126, 1162-1167.
- No Child Left Behind Act (2002). Public Law No. 107-110 107th Congress, 110 *Congressional Record*, 1425, 115 Stat.
- O'Connell, A.A. & Rivet Amico, A. (2010). Logistic regression. In G.R. Hancock & R.O Mueller (Eds.) *The Reviewers Guide to Quantitative Methods in the Social Sciences* pp 221-240.

- O'Connell, A.A., Goldstein, J., Rogers, H.J., & Peng, J.C.Y. (2008). Multilevel logistic models for dichotomous and ordinal data. In A.A. O'Connell & D.B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 199-244).
- Oakes, J.M. (2004). The (mis)estimation of neighborhood effects: causal effects for practical social epidemiology. *Social Science and Medicine*, 54, 1929- 1952.
- Pearl, J. (2010). On a class of bias-amplifying covariates that endanger effect estimates. In P. Grünwald and P. Spirtes (Eds.). *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. AUAI Press. ISBN 978-0-9749039-6-5.
- Pedhazuer, E.J., (1997). *Multiple regression in behavior research: Explanation and prediction*. (3rd Ed.). Australia: Wadsworth.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B* 60(1), 23-40.
- Pruzek, R.M. (2011). Introduction to the special issue on propensity score methods in behavioral research. *Multivariate Behavioral Research*, 46(3), 389-398.
- Quinn, G.P. & Keough, M.J.(2002). *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press
- Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models* (2nd Ed.). Thousand Oaks: Sage.
- Recchia, A. (2010). R-squared measures for two-level hierarchical linear models using SAS. *Journal of Statistical Software*, 32(2),1-9.
- Rodriguez de Gil, P., Lanehart, R., Bellara, A.P., Kromrey, J.D., Kim, E.S., Borman, K.M. & Lee, R. (2012). Propensity score analysis with nested data: Comparing single and multilevel model estimates. In *JSM Proceedings, Social Statistics Section* . Alexandria, VA: American Statistical Association.

- Rosenbaum, P.R. (1986) Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, 11(3), 207-224.
- Rosenbaum, P.R. (1987). Model-based direct adjustment. *The Journal of the American Statistician*, 82, 387-394.
- Rosenbaum, P.R. (2002). *Observational Studies* (2nd Ed.). New York: 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, 516-524.
- Rosenbaum, P.R., & Rubin, D.B. (1985) Constructing a control group using multivariate matching sampling methods that incorporate the propensity score. *American Statistician*, 39, 33-38.
- Rubin, D.B. (1973a). Matching to remove bias in observational studies. *Biometrics* 29(1), 159-183.
- Rubin, D.B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1), 185-203.
- Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D.B. (1976). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, 32(1), 109-120.
- Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34-58.
- Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366), 318-328.

- Rubin, D.B. (1986). Statistics and causal inference: Comment: What ifs have causal answers. *Journal of the American Statistical Association*, 81(396), 961-962.
- Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8), 757-763.
- Rubin, D.B. (2001). Using propensity score to help design observational studies: Application to the Tobacco litigation. *Health Services and Outcomes Research Methodology*, 2, 169-188.
- Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20-30.
- Rubin, D.B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808-840.
- Rubin, D. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, 15(1), 38-46.
- Rubin, D. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, 15(1), 38-46.
- Rubin, D.B. & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52, 249-264.
- Rubin, D.B. & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450), 573-585.
- SAS Institute Inc., SAS® 9.2 Cary, NC: SAS Institute Inc., 2008.
- Scriven, M.J. (2008). A summative evaluation of the RCT methodology: & an alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, 5(9), 11-24.
- Schabenberger, O. (2005). Introducing the GLIMMIX procedure for generalized linear mixed models. *Proceedings from the Annual Meeting of the SAS Users Group International*, 196 (30), 1-20.

- 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. (2000). The empirical program of Quasi-Experimentation. In L. Bickman (Ed.), *Research Design: Donald Campbell's Legacy*. Thousand Oaks, CA: Sage.
- Shadish, W.R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods* 15(1), 3-17.
- Shadish, W.R., & Steiner, P.M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews* 10(1), 19-26
- Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth, Cengage Learning.
- Shah, B.R., Laupacis, A., Hux, J.E., & Austin, P.C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, 58, 550-559.
- Slavin, R. E. (2002) Evidence-based educational policies: Transforming educational practice and research. *Educational Researcher*, 31, 15–21
- Slavin, R.E. (2008) Perspectives on evidenced based research in education —what works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1),5-14.
- Smith, J.A., & Todd, P.E. (2001) Reconciling conflicting evidence on the performance of propensity score matching methods. *The American Economic Review* 91(2), 112-118.
- Smith, J.A., & Todd, P.E. (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125, 305-353.
- Steiner, P.M. & Cook, T.D. (2013). Matching and Propensity Scores. In T.D. Little (Ed.). *The Oxford Handbook of Quantitative Methods Vol.1*. Cary, NC: Oxford University Press.

- Steiner, P. M., Cook, T.D., & Shadish, W.R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, 36(2), 213-236.
- Steiner, P. M., Cook, T.D., Shadish, W.R, and Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3), 250-267.
- Stuart, E.A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher*, 36(4), 187-198.
- Stuart, E.A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996-2003' by Peter Austin, *Statistics in Medicine*. *Statistics in Medicine*, 27, 2062-2065.
- Stuart, E.A. (2010). Matching methods for causal inference: A review and look forward. *Statistical Science*, 25(1), 1-21.
- Stuart, E.A. & Green, K.M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44(2), 395-406.
- Sturmer, T., Joshi, M.M., Glynn, R.J., Avron, J., Rothman, K., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different methods compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5) 437-447.
- Thoemmes, F.J. (2009) The use of propensity scores with clustered data: A simulation study. Unpublished Dissertation. AAT: 3380671.
- Thoemmes, F., & Kim, E.S., (2011). A Systematic Review of Propensity Score Methods in the Social Sciences', *Multivariate Behavioral Research*, 46 (1), 90 -118.

- Thoemmes, F., & West, S.G. (2011) The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46(3), 514-543.
- U.S. Department of Education. (2003, November 4). Scientifically based evaluation methods. Federal Registrar, 62445-62447.
- U.S. Department of Education. (2005, November 4). Scientifically based evaluation methods. Federal Registrar, 62445-62447.
- 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. *Pharmacoepidmiology and Drug Safety*, 13, 841-853.
- Weitzen, S., Lapane, K.L., Toledano, A.Y., Hume, A.L., & Mor, V. (2005). Weaknesses of goodness-of-fit test for evaluating propensity score models: The case of the omitted confounder. *Pharmacoepidmiology and Drug Safety*, 14, 227-238.
- West, S.G., & Thoemmes, F., (2010). Campbell's and Rubin's perspectives on Causal Inference. *Psychological Methods*, 15(1), 18-37.
- Wilkinson, L. & Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* 54(8), 594-604.
- Wooldridge, J.M. (2009). Should instrumental variables be used as matching variables? Tech Rep. Retrieved from:
<https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf>, Michigan State University, MI.

APPENDIX A

Equation for Data Generation

$$\begin{aligned} Y_{ij} = & \beta_{0j} + \beta_{1j}X1_{ij} + \beta_{2j}X2_{ij} + \beta_{3j}X3_{ij} + \beta_{4j}X4_{ij} + \beta_{5j}X5_{ij} + \beta_{6j}X6_{ij} + \\ & \beta_{7j}X7_{ij} + \beta_{8j}X8_{ij} + \beta_{9j}X9_{ij} + \beta_{10j}X10_{ij} + \beta_{11j}X11_{ij} + \beta_{12j}X12_{ij} + \beta_{13j}X13_{ij} + \\ & \beta_{14j}X14_{ij} + \beta_{15j}X15_{ij} + \beta_{16j}X16_{ij} + \beta_{17j}X17_{ij} + \beta_{18j}X18_{ij} + \beta_{19j}X19_{ij} + \\ & \beta_{20j}X20_{ij} + \beta_{21j}X21_{ij} + \beta_{22j}X22_{ij} + \beta_{23j}X23_{ij} + \beta_{24j}X24_{ij} + \beta_{25j}X25_{ij} + \\ & \beta_{26j}X26_{ij} + \beta_{27j}X27_{ij} + \beta_{28j}X28_{ij} + \beta_{29j}X29_{ij} + \beta_{30j}X30_{ij} + \beta_{zj}Z_{ij} + e_{ij} \end{aligned}$$

Appendix A (Continued)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

$$\beta_{2j} = \gamma_{20} + \mu_{2j}$$

$$\beta_{3j} = \gamma_{30} + \mu_{3j}$$

$$\beta_{4j} = \gamma_{40} + \mu_{4j}$$

$$\beta_{5j} = \gamma_{50} + \mu_{5j}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80} + \gamma_{81}W1_j$$

$$\beta_{9j} = \gamma_{90} + \gamma_{91}W1_j$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}W1_j$$

$$\beta_{11j} = \gamma_{110}$$

$$\beta_{12j} = \gamma_{120}$$

$$\beta_{13j} = \gamma_{130}$$

$$\beta_{14j} = \gamma_{140}$$

$$\beta_{15j} = \gamma_{150}$$

$$\beta_{16j} = \gamma_{160}$$

$$\beta_{17j} = \gamma_{170}$$

$$\beta_{18j} = \gamma_{180}$$

$$\beta_{19j} = \gamma_{190}$$

$$\beta_{20j} = \gamma_{200}$$

$$\beta_{21j} = \gamma_{210}$$

$$\beta_{22j} = \gamma_{220}$$

$$\beta_{23j} = \gamma_{230}$$

$$\beta_{24j} = \gamma_{240}$$

$$\beta_{25j} = \gamma_{250}$$

$$\beta_{26j} = \gamma_{260}$$

$$\beta_{27j} = \gamma_{270}$$

$$\beta_{28j} = \gamma_{280}$$

$$\beta_{29j} = \gamma_{290}$$

$$\beta_{30j} = \gamma_{300}$$

$$\beta_{Zj} = \gamma_{Z0}$$

Appendix A (Continued)

Combined as,

$$\begin{aligned}
 Y_{ij} = & \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \\
 & \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \gamma_{10}X1_{ij} + \gamma_{20}X2_{ij} + \gamma_{30}X3_{ij} + \gamma_{40}X4_{ij} + \gamma_{50}X5_{ij} + \\
 & \gamma_{60}X6_{ij} + \gamma_{70}X7_{ij} + \gamma_{80}X8_{ij} + \gamma_{81}W1_j X8_{ij} + \gamma_{90}X9_{ij} + \gamma_{91}W1_j X9_{ij} + \gamma_{100}X10_{ij} + \\
 & \gamma_{101}W1_j X10_{ij} + \gamma_{110}X11_{ij} + \gamma_{120}X12_{ij} + \gamma_{130}X13_{ij} + \gamma_{140}X14_{ij} + \gamma_{150}X15_{ij} + \gamma_{160}X16_{ij} + \\
 & \gamma_{170}X17_{ij} + \gamma_{180}X18_{ij} + \gamma_{190}X19_{ij} + \gamma_{200}X20_{ij} + \gamma_{210}X21_{ij} + \gamma_{220}X22_{ij} + \gamma_{230}X23_{ij} + \\
 & \gamma_{240}X24_{ij} + \gamma_{250}X25_{ij} + \gamma_{260}X26_{ij} + \gamma_{270}X27_{ij} + \gamma_{280}X28_{ij} + \gamma_{290}X29_{ij} + \gamma_{300}X30_{ij} + \gamma_{z0}Z_{ij} \\
 & \mu_{0j} + \mu_{1j}X1_{ij} + \mu_{2j}X2_{ij} + \mu_{3j}X3_{ij} + \mu_{4j}X4_{ij} + \mu_{5j}X5_{ij} \\
 e_{ij} \sim & N(0, \sigma^2)
 \end{aligned}$$

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ & \tau_{11} & & & & \\ & & \tau_{22} & & & \\ & & & \tau_{33} & & \\ & & & & \tau_{44} & \\ & & & & & \tau_{55} \end{pmatrix} \right)$$

APPENDIX B

Population R² Values Simulated

Table A1

R² values for the different confounder magnitudes

	$\delta=0.2$	$\delta=0.5$
SSSS	.1425332	.2779931
SSSM	.2500427	.3667704
SSMS	.4831581	.4899146
SSMM	.6020114	.5890356
SMSS	.1425197	.2695686
SMSM	.2585595	.3455621
SMMS	.4903178	.4878974
SMMM	.6266342	.5841673
MMMM	.75	.5768803
MMMS	.5951017	.3568233
MMSM	.27	.3568233
MMSS	.1488168	.267808
MSSS	.1446851	.2765483
MSMM	.715	.5691231
MSMS	.5870403	.4603783
MSSM	.2585468	.3711888

Note: S denotes small relationship and M denotes moderate relationship

The order of the combinations=XZ, WZ, XY, WY

Small XZ and XY=.10; Moderate XZ and XY=.20

Small WZ and WY=.20; Moderate WZ and WY=.40

APPENDIX C

Equations for Propensity Score Estimation

Single Level Model

$$\widehat{\text{logitPS}} = \beta_{00} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16} + \beta_{17} X_{17} + \beta_{18} X_{18} + \beta_{19} X_{19} + \beta_{20} X_{20} + \beta_{21} X_{21} + \beta_{22} X_{22} + \beta_{23} X_{23} + \beta_{24} X_{24} + \beta_{25} X_{25} + \beta_{26} X_{26} + \beta_{27} X_{27} + \beta_{28} X_{28} + \beta_{29} X_{29} + \beta_{30} X_{30} + \beta_{31} W_1 + \beta_{32} W_2 + \beta_{33} W_3 + \beta_{34} W_4 + \beta_{35} W_5 + \beta_{36} W_6 + \beta_{37} W_7 + \beta_{38} W_8 + \beta_{39} W_9 + \beta_{40} W_{10}$$

Random Intercepts Model

$$\widehat{\text{logit PS}}_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \beta_{5j} X_{5ij} + \beta_{6j} X_{6ij} + \beta_{7j} X_{7ij} + \beta_{8j} X_{8ij} + \beta_{9j} X_{9ij} + \beta_{10j} X_{10ij} + \beta_{11j} X_{11ij} + \beta_{12j} X_{12ij} + \beta_{13j} X_{13ij} + \beta_{14j} X_{14ij} + \beta_{15j} X_{15ij} + \beta_{16j} X_{16ij} + \beta_{17j} X_{17ij} + \beta_{18j} X_{18ij} + \beta_{19j} X_{19ij} + \beta_{20j} X_{20ij} + \beta_{21j} X_{21ij} + \beta_{22j} X_{22ij} + \beta_{23j} X_{23ij} + \beta_{24j} X_{24ij} + \beta_{25j} X_{25ij} + \beta_{26j} X_{26ij} + \beta_{27j} X_{27ij} + \beta_{28j} X_{28ij} + \beta_{29j} X_{29ij} + \beta_{30j} X_{30ij}$$

APPENDIX C (Continued)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

$$\beta_{10j} = \gamma_{100}$$

$$\beta_{11j} = \gamma_{110}$$

$$\beta_{12j} = \gamma_{120}$$

$$\beta_{13j} = \gamma_{130}$$

$$\beta_{14j} = \gamma_{140}$$

$$\beta_{15j} = \gamma_{150}$$

$$\beta_{16j} = \gamma_{160}$$

$$\beta_{17j} = \gamma_{170}$$

$$\beta_{18j} = \gamma_{180}$$

$$\beta_{19j} = \gamma_{190}$$

$$\beta_{20j} = \gamma_{200}$$

$$\beta_{21j} = \gamma_{210}$$

$$\beta_{22j} = \gamma_{220}$$

$$\beta_{23j} = \gamma_{230}$$

$$\beta_{24j} = \gamma_{240}$$

$$\beta_{25j} = \gamma_{250}$$

$$\beta_{26j} = \gamma_{260}$$

$$\beta_{27j} = \gamma_{270}$$

$$\beta_{28j} = \gamma_{280}$$

$$\beta_{29j} = \gamma_{290}$$

$$\beta_{30j} = \gamma_{300}$$

APPENDIX C (Continued)

Combined as:

$$\begin{aligned} \text{logit PS}_{ij} = & \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \\ & \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \gamma_{10}X1_{ij} + \gamma_{20}X2_{ij} + \gamma_{30}X3_{ij} + \gamma_{40}X4_{ij} + \gamma_{50}X5_{ij} + \\ & \gamma_{60}X6_{ij} + \gamma_{70}X7_{ij} + \gamma_{80}X8_{ij} + \gamma_{90}X9_{ij} + \gamma_{100}X10_{ij} + \gamma_{110}X11_{ij} + \gamma_{120}X12_{ij} + \\ & \gamma_{130}X13_{ij} + \gamma_{140}X14_{ij} + \gamma_{150}X15_{ij} + \gamma_{160}X16_{ij} + \gamma_{170}X17_{ij} + \gamma_{180}X18_{ij} + \gamma_{190}X19_{ij} + \\ & \gamma_{200}X20_{ij} + \gamma_{210}X21_{ij} + \gamma_{220}X22_{ij} + \gamma_{230}X23_{ij} + \gamma_{240}X24_{ij} + \gamma_{250}X25_{ij} + \gamma_{260}X26_{ij} + \\ & \gamma_{270}X27_{ij} + \gamma_{280}X28_{ij} + \gamma_{290}X29_{ij} + \gamma_{300}X30_{ij} + \mu_{0j} \\ & \mu_{0j} \sim N(0, \tau_{00}) \end{aligned}$$

Random Coefficients Model

$$\begin{aligned} \text{logitPS}_{ij} = & \beta_{00} + \beta_1X_{1ij} + \beta_2X_{2ij} + \beta_3X_{3ij} + \beta_4X_{4ij} + \beta_5X_{5ij} + \beta_6X_{6ij} + \beta_7X_{7ij} + \beta_8X_{8ij} + \\ & \beta_9X_{9ij} + \beta_{10}X_{10ij} + \beta_{11}X_{11ij} + \beta_{12}X_{12ij} + \beta_{13}X_{13ij} + \beta_{14}X_{14ij} + \beta_{15}X_{15ij} + \beta_{16}X_{16ij} + \beta_{17}X_{17ij} + \\ & \beta_{18}X_{18ij} + \beta_{19}X_{19ij} + \beta_{20}X_{20ij} + \beta_{21}X_{21ij} + \beta_{22}X_{22ij} + \beta_{23}X_{23ij} + \beta_{24}X_{24ij} + \beta_{25}X_{25ij} + \\ & \beta_{26}X_{26ij} + \beta_{27}X_{27ij} + \beta_{28}X_{28ij} + \beta_{29}X_{29ij} + \beta_{30}X_{30ij} \end{aligned}$$

APPENDIX C (Continued)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

$$\beta_{2j} = \gamma_{20} + \mu_{2j}$$

$$\beta_{3j} = \gamma_{30} + \mu_{3j}$$

$$\beta_{4j} = \gamma_{40} + \mu_{4j}$$

$$\beta_{5j} = \gamma_{50} + \mu_{5j}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

$$\beta_{10j} = \gamma_{100}$$

$$\beta_{11j} = \gamma_{110}$$

$$\beta_{12j} = \gamma_{120}$$

$$\beta_{13j} = \gamma_{130}$$

$$\beta_{14j} = \gamma_{140}$$

$$\beta_{15j} = \gamma_{150}$$

$$\beta_{16j} = \gamma_{160}$$

$$\beta_{17j} = \gamma_{170}$$

$$\beta_{18j} = \gamma_{180}$$

$$\beta_{19j} = \gamma_{190}$$

$$\beta_{20j} = \gamma_{200}$$

$$\beta_{21j} = \gamma_{210}$$

$$\beta_{22j} = \gamma_{220}$$

$$\beta_{23j} = \gamma_{230}$$

$$\beta_{24j} = \gamma_{240}$$

$$\beta_{25j} = \gamma_{250}$$

$$\beta_{26j} = \gamma_{260}$$

$$\beta_{27j} = \gamma_{270}$$

$$\beta_{28j} = \gamma_{280}$$

$$\beta_{29j} = \gamma_{290}$$

$$\beta_{30j} = \gamma_{300}$$

APPENDIX C (Continued)

Combined as:

$$\begin{aligned} \text{logit PS}_{ij} = & \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \\ & \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \gamma_{10}X1_{ij} + \gamma_{20}X2_{ij} + \gamma_{30}X3_{ij} + \gamma_{40}X4_{ij} + \gamma_{50}X5_{ij} + \\ & \gamma_{60}X6_{ij} + \gamma_{70}X7_{ij} + \gamma_{80}X8_{ij} + \gamma_{90}X9_{ij} + \gamma_{100}X10_{ij} + \gamma_{110}X11_{ij} + \gamma_{120}X12_{ij} + \\ & \gamma_{130}X13_{ij} + \gamma_{140}X14_{ij} + \gamma_{150}X15_{ij} + \gamma_{160}X16_{ij} + \gamma_{170}X17_{ij} + \gamma_{180}X18_{ij} + \gamma_{190}X19_{ij} + \\ & \gamma_{200}X20_{ij} + \gamma_{210}X21_{ij} + \gamma_{220}X22_{ij} + \gamma_{230}X23_{ij} + \gamma_{240}X24_{ij} + \gamma_{250}X25_{ij} + \gamma_{260}X26_{ij} + \\ & \gamma_{270}X27_{ij} + \gamma_{280}X28_{ij} + \gamma_{290}X29_{ij} + \gamma_{300}X30_{ij} + \\ & \mu_{0j} + \mu_{1j}X1_{ij} + \mu_{2j}X2_{ij} + \mu_{3j}X3_{ij} + \mu_{4j}X4_{ij} + \mu_{5j}X5_{ij} \end{aligned}$$

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ & \tau_{11} & & & & \\ & & \tau_{22} & & & \\ & & & \tau_{33} & & \\ & & & & \tau_{44} & \\ & & & & & \tau_{55} \end{pmatrix} \right)$$

Cross level model

$$\begin{aligned} \text{logitPS}_{ij} = & \beta_{00} + \beta_1X1_{ij} + \beta_2X2_{ij} + \beta_3X3_{ij} + \beta_4X4_{ij} + \beta_5X5_{ij} + \beta_6X6_{ij} + \beta_7X7_{ij} + \beta_8X8_{ij} + \\ & \beta_9X9_{ij} + \beta_{10}X10_{ij} + \beta_{11}X11_{ij} + \beta_{12}X12_{ij} + \beta_{13}X13_{ij} + \beta_{14}X14_{ij} + \beta_{15}X15_{ij} + \beta_{16}X16_{ij} + \beta_{17}X17_{ij} + \\ & \beta_{18}X18_{ij} + \beta_{19}X19_{ij} + \beta_{20}X20_{ij} + \beta_{21}X21_{ij} + \beta_{22}X22_{ij} + \beta_{23}X23_{ij} + \beta_{24}X24_{ij} + \beta_{25}X25_{ij} + \\ & \beta_{26}X26_{ij} + \beta_{27}X27_{ij} + \beta_{28}X28_{ij} + \beta_{29}X29_{ij} + \beta_{30}X30_{ij} \end{aligned}$$

APPENDIX C (Continued)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \mu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

$$\beta_{2j} = \gamma_{20} + \mu_{2j}$$

$$\beta_{3j} = \gamma_{30} + \mu_{3j}$$

$$\beta_{4j} = \gamma_{40} + \mu_{4j}$$

$$\beta_{5j} = \gamma_{50} + \mu_{5j}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80} + \gamma_{81}W1_j$$

$$\beta_{9j} = \gamma_{90} + \gamma_{91}W1_j$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}W1_j$$

$$\beta_{11j} = \gamma_{110}$$

$$\beta_{12j} = \gamma_{120}$$

$$\beta_{13j} = \gamma_{130}$$

$$\beta_{14j} = \gamma_{140}$$

$$\beta_{15j} = \gamma_{150}$$

$$\beta_{16j} = \gamma_{160}$$

$$\beta_{17j} = \gamma_{170}$$

$$\beta_{18j} = \gamma_{180}$$

$$\beta_{19j} = \gamma_{190}$$

$$\beta_{20j} = \gamma_{200}$$

$$\beta_{21j} = \gamma_{210}$$

$$\beta_{22j} = \gamma_{220}$$

$$\beta_{23j} = \gamma_{230}$$

$$\beta_{24j} = \gamma_{240}$$

$$\beta_{25j} = \gamma_{250}$$

$$\beta_{26j} = \gamma_{260}$$

$$\beta_{27j} = \gamma_{270}$$

$$\beta_{28j} = \gamma_{280}$$

$$\beta_{29j} = \gamma_{290}$$

$$\beta_{30j} = \gamma_{300}$$

APPENDIX C (Continued)

Combined as,

$$\begin{aligned} \text{logit PS}_{ij} = & \gamma_{00} + \gamma_{01}W1_j + \gamma_{02}W2_j + \gamma_{03}W3_j + \gamma_{04}W4_j + \gamma_{05}W5_j + \gamma_{06}W6_j + \gamma_{07}W7_j + \\ & \gamma_{08}W8_j + \gamma_{09}W9_j + \gamma_{010}W10_j + \gamma_{10}X1_{ij} + \gamma_{20}X2_{ij} + \gamma_{30}X3_{ij} + \gamma_{40}X4_{ij} + \gamma_{50}X5_{ij} + \\ & \gamma_{60}X6_{ij} + \gamma_{70}X7_{ij} + \gamma_{80}X8_{ij} + \gamma_{81}W1_j X8_{ij} + \gamma_{90}X9_{ij} + \gamma_{91}W1_j X9_{ij} + \gamma_{100}X10_{ij} + \\ & \gamma_{101}W1_j X10_{ij} + \gamma_{110}X11_{ij} + \gamma_{120}X12_{ij} + \gamma_{130}X13_{ij} + \gamma_{140}X14_{ij} + \gamma_{150}X15_{ij} + \gamma_{160}X16_{ij} + \\ & \gamma_{170}X17_{ij} + \gamma_{180}X18_{ij} + \gamma_{190}X19_{ij} + \gamma_{200}X20_{ij} + \gamma_{210}X21_{ij} + \gamma_{220}X22_{ij} + \gamma_{230}X23_{ij} + \\ & \gamma_{240}X24_{ij} + \gamma_{250}X25_{ij} + \gamma_{260}X26_{ij} + \gamma_{270}X27_{ij} + \gamma_{280}X28_{ij} + \gamma_{290}X29_{ij} + \gamma_{300}X30_{ij} + \\ & \mu_{0j} + \mu_{1j}X1_{ij} + \mu_{2j}X2_{ij} + \mu_{3j}X3_{ij} + \mu_{4j}X4_{ij} + \mu_{5j}X5_{ij} \end{aligned}$$

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ & \tau_{11} & & & & \\ & & \tau_{22} & & & \\ & & & \tau_{33} & & \\ & & & & \tau_{44} & \\ & & & & & \tau_{55} \end{pmatrix} \right)$$