EXPLORING THE JOINT USE OF PROPENSITY AND PROGNOSTIC SCORES FOR BIAS REDUCTION IN THE CONTEXT OF SMALL EDUCATIONAL PROGRAM EVALUATIONS

by

Yun Tang

Beihang University, B.A., 2000

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Yun Tang

It was defended on

February 27, 2018

and approved by

Feifei Ye, Ph.D., Senior Behavioral/Social Scientist, RAND Corporation

Lindsay Page, Ph.D., Assistant Professor, Department of Psychology in Education

Fengyan Tang, Ph.D., Professor, School of Social Work

Dissertation Advisor: Clement Stone, Ph.D., Professor, Department of Psychology in Education

EXPLORING THE JOINT USE OF PROPENSITY AND PROGNOSTIC SCORES FOR BIAS REDUCTION IN THE CONTEXT OF SMALL EDUCATIONAL PROGRAM EVALUATIONS

Yun Tang, Ph.D.

University of Pittsburgh, 2018

Propensity and prognostic score methods are two statistical techniques used to correct for the selection bias in nonexperimental studies. Recently, the joint use of propensity and prognostic scores (i.e., two-score methods) has been proposed to improve the performance of adjustments using propensity or prognostic scores alone for bias reduction. The main purpose of this dissertation study was to evaluate the effectiveness of the joint use of propensity and prognostic scores for reducing bias of treatment effect estimates in quasi-experimental designs. To this end, a simulation study based on real educational data was conducted to investigate the comparative performance of separate and combined use of propensity and prognostic scores for recovering a simulated treatment effect under various conditions. These conditions were based on different control group sizes, outcome measures, and propensity score estimation methods. Specifically, four two-score methods were examined in this study: weighting and 1:1 optimal matching on the estimated prognostic propensity scores, and 1:1 and full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores. Single score adjustments that were examined included 1:1 matching on the estimated propensity or prognostic scores, and weighting on the estimated propensity scores. The simulation results did not support the use of any of the

two-score methods as alternatives to single score adjustments in estimation of treatment effects in the examined conditions. Instead, matching on the estimated prognostic scores showed some advantages over all the two-score methods and single score adjustments involving propensity scores only. However, this seemingly promising finding for adjustments on prognostic scores is tempered by the inherent “in-sample” problems for estimating prognostic scores.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EQUATIONS

PREFACE

The amazing road that led me to complete this dissertation was a long, but very rewarding experience. During this lengthy and challenging journey, the person I want to thank first and foremost is my doctoral advisor, Professor Clement Stone. I am deeply grateful for the tremendous supervision, encouragement, and support I received from him over these years. This dissertation would not have been possible without his brilliant ideas, instructive suggestions, and constructive criticisms. His guidance became the main driving force behind the entire project, and it is hard to put my deep appreciation into adequate words here. I am also extremely grateful to my committee member, Dr. Feifei Ye, not only for her guidance and support, but also for her encouragement and belief in my work during those times when I doubted my capabilities. I also want to extend my sincerest appreciation to my other committee members, Professor Lindsay Page and Professor Fengyan Tang, for their kind advice and support during the various stages of this project. A special thanks also goes to Professor Yu (Joyce) Jiang at the University of Memphis, for not only helping me acquire the data I needed to write my dissertation, but also for the helpful discussions offered on my research and the ongoing advice on my career path. I would also like to express my gratitude to Professor Suzanne Lane and the late Professor Kevin Kim for the courses they taught and for being extremely supportive whenever needed. I owe an immense amount of gratitude to the Research Methodology program as a whole for giving me the opportunity to receive a first-class education in an area about which I am passionate and truly inspired to give ever more to in the years ahead.

I further want to acknowledge the company of my many graduate fellows and friends at PITT – Hong, Jie, Yi, Liqun, Ting, Meisian, Caiyan, Meng, Xiaoyan – and also the many other individuals, in no particular order. I will never forget the conversations, laughter, and meals we have shared. Thank you all for truly making Pittsburgh my home away from home.

I am also thankful to my colleagues Brenda, Sheila, and Candy at CREP at the University of Memphis for their friendship, encouragement, and belief in me. A special thanks also goes to my supervisor, Dr. Todd Zoblotsky at CREP for being supportive and giving me the time I needed to be able to complete my dissertation project successfully.

Finally, my greatest appreciation goes to my dear husband, Sergio, and my beloved parents, for their unconditional love, support, and patience. Thank you for being there. My education and accomplishments are forever offered in loving dedication to each of them.

1.0 INTRODUCTION

1.1 BACKGROUND

Randomized controlled trials (RCTs) have long been considered as the "gold standard" in causal inference research (Shadish, Cook, & Campbell, 2002). Random assignment ensures that, on average, both treatment and control groups are equivalent in terms of both observed and unobserved characteristics prior to intervention. Thus, the observed differences between treatment and control groups are more likely due to the treatment than to their preexisting group differences. That is why researchers prefer utilizing an RCT for impact evaluations whenever it is possible.

However, randomization is often unethical and/or infeasible in many educational settings. In these cases, researchers in education have to rely on quasi-experimental (QE) designs to investigate causal questions (Cook, 2002; Shadish et al., 2002). In fact, RCTs are underrepresented in educational evaluations. Among studies reviewed by What Works Clearinghouse, the most common method for evaluating program impacts was non-equivalent group observational[1] studies; only 30% were RCTs (Wong, Valentine, & Miller-Bains, 2017). In QE designs, treatment and control groups are not created by randomization. Instead, research subjects self-select or are selected into an intervention group by a third-party. As a result, treatment and control groups in observational studies may have systematic preexisting differences. These systematic differences

[1] Throughout my dissertation, I use "QE" and "observational study" interchangeably.

can confound with treatment and thus lead to biased inferences about treatment effects. This threat to the internal validity of the conclusion is called selection bias. The existence of selection bias makes it difficult to determine causal relationships in observational studies.

A number of methods have been developed to adjust for selection bias in observational studies. Conventional multivariable regression model, which includes the confounding variables as the explanatory variables, is one of the most frequently used adjustment methods. However, the regression estimates can be severely biased when the model is misspecified, or if the treatment and control groups differ greatly in observed characteristics (Rubin, 1997, 2001; Schafer & Kang, 2008).

Another frequently used method is exact matching. This method matches subjects based on exact values of background covariates. Compared to regression, matching does not rely on the assumptions of functional form. However, it has a dimensionality problem. As the number of matching variable gets large, the matching can be very difficult or even impossible. And the exact matching is even more difficult for continuous covariates. In addition to exact matching, multivariate matching methods, such as Mahalanobis distance matching, also has a dimensionality problem (Rubin, 1979; Gu & Rosenbaum, 1993). That is, Mahalanobis distance is not good for matching that involves a large number of matching covariates. Therefore, methods that combine the information from multiple confounding variables into a summary score are more desirable.

Rosenbaum and Rubin (1983, 1984, 1985) introduced one statistical adjustment method called propensity score (PS) to solve this dimensionality problem. Research has shown that PS works better than Mahalanobis distance matching when the number of covariates is larger than five (Gu & Rosenbaum, 1993; Rubin & Thomas, 2000). A propensity score, defined as the "conditional probability of assignment to a particular treatment given a vector of observed

covariates" (Rosenbaum and Rubin, 1983, p. 41), involves a data reduction technique that summarizes a vector of observed covariates into one single score. Propensity scores are often estimated by logistic regression (D'Agostino, 1998). Once the propensity score is estimated, researchers can use it for matching, stratification, weighting, or covariance adjustment to balance two non-equivalent groups on observed covariates and consequently obtain more accurate estimates of treatment effects (Schafer & Kang, 2008; Stuart, 2010).

Another summary score is the prognostic score (Hansen, 2008). Prognostic scores (PROGs) are defined as any scalar or multidimensional function of the covariates that, when conditioned on, results in covariates being independent of the potential outcomes under the control condition (Hansen, 2008). Unlike PSs which model the associations between treatment assignment and the observed covariates, PROGs model the relationships between potential outcomes in the control condition and the observed covariates. Hansen (2008) argued that, given PROGs, the difference across comparison groups can be attributed to the treatment since the relationship between outcomes and covariates have been controlled. Thus, the estimated PROGs can be used to adjust for selection bias via methods similar to those used with PSs (Hansen, 2008).

Since its introduction in 1983, PS analysis has been extensively researched. Adjustments using PSs have been effective in reducing or eliminating selection bias when properly used and they have replicated results from RCTs under certain conditions (e.g., Luellen, Shadish, & Clark, 2005; Shadish, Clark, & Steiner, 2008). As a result, the application of PS analysis in the social and educational fields has become popular over the past 10 years (Thoemmes & Kim, 2011; Hernandez, 2015). Applied researchers have used PS methods to look at effects of various intervention programs in many educational settings (e.g., K-12 and higher education). The examined interventions include, but are not limited to, dual enrollment (An, 2012), distance

education (Xu & Jaggars, 2011), kindergarten retention policy (Hong & Raudenbush, 2005, 2006), Catholic versus public schools (Morgan, 2001), STEM (Wang, 2014), and special education (Sullivan & Field, 2013).

In contrast, very few studies have examined the performance of PROGs for bias reduction. Moreover, the application of PROGs for confounding control in educational research is extremely limited. To the best of my knowledge, only two case studies have used PROGs to control for remaining imbalances either in an RCT study (Pane, Griffin, McCaffrey, & Karam, 2014) or in a QE study (Garret & Hong, 2016). When estimating the program impact, Pane et al. (2014) adjusted for both pretreatment covariates and PROGs in the outcome model, and Garret and Hong (2016) used PROGs as the sole predictor in the outcome model within the strata created through a PS adjustment.

Although scholars have found PROGs to be an effective alternative to PSs in certain settings (e.g., Arbogast & Ray, 2011; Stürmer et al., 2005), Hansen (2008) viewed PROG as a complement rather than an alternative to PS due to some inherent problems with the estimation of PROGs. Thus, Hansen (2008) proposed a new QE approach that involved the joint use of PSs and PROGs (i.e., the two-score method). Hansen argued that, in theory, using both scores may be preferable to using a PS adjustment with regard to reducing bias and/or improving the precision of the treatment effect estimates. To varying degrees, two simulation studies (Hansen, 2006; Leacy & Stuart, 2014) have found evidence to support Hansen's hypothesis. Despite the potential for jointly using PSs and PROGs in bias reduction, the research on this new QE method has been very limited and restricted to unrealistic settings. Hence, more research is needed to identify the conditions under which the two-score method could work.

1.2 PURPOSE OF THE STUDY

The primary purpose of this study was to investigate whether the joint use of propensity and prognostic scores was more effective than the use of propensity or prognostic scores alone for reducing bias of treatment effect estimates in QE studies. To this end, a simulation study in an educational context was conducted to compare the relative performance of two-score methods and single score methods for recovering a simulated treatment effect under experimental conditions based on different control group sizes, outcome measures, and PS estimation methods. It is hoped that the findings from this study could inform educational researchers' design decisions when they evaluate impact studies using observational data.

1.3 ORGANIZATION OF THE DISSERTATION

The organization of this dissertation study is as follows. Chapter 1 is an introduction to the study's background and purpose. Chapter 2 is a review of the existing literature on PSs, PROGs, and the joint use of those scores. In Chapter 3, the simulation design of the study is explained in detail. The results of the simulation study are presented in Chapter 4. Finally, Chapter 5 includes a discussion of the simulation results and their implications for impact evaluations in the field of education. I also discuss this study's limitations and directions for future research in this chapter.

2.0 LITERATURE REVIEW

Researchers often rely on quasi-experimental (QE) methods to draw causal inferences when randomized controlled trials (RCTs) are impossible for ethical or practical reasons. The current study focuses on the QE methods that use propensity scores (PSs), prognostic scores (PROGs), and a combination of both scores. Rubin's (1974) causal model is fundamental to an understanding of the propensity and prognostic scores and to causal inference.

2.1 RUBIN'S CAUSAL MODEL

Propensity score (PS) analysis is developed in the framework of the Rubin's causal model, also known as the Neyman-Rubin potential outcomes model (Rubin, 1974, 2005). Under this framework for a binary treatment, each subject *i* has two potential outcomes, one corresponding to the treatment condition (Y_{i1}), and the other corresponding to the control condition (Y_{i0}). Thus, the potential outcomes for the unit i can be expressed as $Y_i = Z_i Y_{i1} - (1 - Z_i) Y_{i0}$, where Z_i indicates unit *i*'s treatment condition ($Z_i = 1$ for treatment; $Z_i = 0$ for control).

Suppose both potential outcomes can be observed at the same time. The causal effect for subject *i* is defined as the difference between these two potential outcomes, $\delta_i = Y_{i1} - Y_{i0}$ (Rubin, 1974). However, both potential outcomes for a subject cannot be observed at the same time. Only the outcome under one condition can be observed in practice. This is what Holland refers to as the

"fundamental problem of causal inference" (Holland, 1986, p. 947). Specifically, treated subjects only have potential outcomes in the treatment condition; their potential outcomes in the control condition are missing. For control subjects, their potential outcomes in the control condition are available, while their treatment potential outcomes are missing. Therefore, it is impossible to estimate the causal effect for an individual subject. Instead, only an average effect of a group of subjects can be estimated. In order to estimate average treatment effect, one has to rely on the group averages of observed outcomes to predict the counterfactuals (i.e., unobserved potential outcomes) of the treatment subjects in the control condition, and/or the counterfactuals of the control subjects in the treatment condition, depending on the type of treatment effect that is of interest.

2.1.1 Type of Treatment Effect

There are different types of average treatment effects discussed in the causal inference literature (see Guo & Fraser, 2010). The effects relevant to PS analysis are the average treatment effect (ATE) and the average treatment effect on the treated (ATT). ATE is considered as the effect of a program or intervention on the entire population: $\delta = E[(Y_1|Z = 1) - (Y_0|Z = 0)]$. ATT aims to investigate the effect of a program or intervention on the research subjects who receive or would receive the treatment, $\delta = E[(Y_1 - Y_0)|Z = 1]$. From the perspective of program evaluation, ATE assesses whether on average the program is beneficial for all individuals, while ATT evaluates whether on average the program is beneficial on individuals who actually participate in the program. In practice, ATT, rather than ATE, is more of substantive interest, as ATT provides information regarding whether to continue policies or programs that target a specific

group of people who are most likely to benefit from the intended policies or programs (Schafer & Kang, 2008).

2.2 PROPENSITY SCORE ADJUSTMENT

The basic idea of PS methods is to derive a scalar based on covariates that account for any observed pretreatment group differences. The scalar variable is called the PS and it is defined as the conditional probability of assignment to a treatment given a vector of observed covariates (Rosenbaum & Rubin, 1983, p. 42):

$$e(x) = pr(z = 1|x) \quad (1)$$

where $e(x)$ is the estimated PS, z indicates the treatment assignment, and x is a vector of observed covariates.

There are two underlying assumptions for PSs (Rosenbaum & Rubin, 1983). The first assumption is the strongly ignorable treatment assignment assumption. This assumption consists of two components. First, it suggests that the potential outcomes are independent of the treatment assignment given the observed covariates: $(Y_0, Y_1) \perp Z|X$. This assumption is satisfied if all the covariates that affect the treatment assignment have been accounted for, so that there are no unobserved covariates that will affect the effect estimation (i.e., no hidden bias). Another important component of this assumption is the common support assumption: $0 < pr(Z = 1|X) < 1$. This assumption suggests that every unit has a probability of being assigned to treatment, and there is an overlap in terms of PSs between treatment and control groups. If this assumption is not satisfied, it indicates treatment and control groups are very different groups, and no valid inferences can be drawn from the comparison. The estimate is merely a result of extrapolation.

The second fundamental assumption underlying PS is the stable unit treatment value assumption (SUTVA) (Rubin, 1980, 1986). This assumption has two components. The first is that the mechanism for assigning participants to treatment and control groups should not affect their responses. The second is that the potential outcomes for one participant is not affected by the treatment received by other participants. This assumption incorporates the ideas that units do not interfere with each other and that for each subject there is only one single version of each treatment level. This assumption allows us to model the outcome of one unit independent of another unit's treatment status, given the observed covariates.

An important feature of PS is its balancing property. PS is a balancing score such that the conditional distribution of the pretreatment covariates, given the PS, is the same between the treatment and control groups (Rosenbaum & Rubin, 1983). In other words, if the treated and control subjects have the same PSs, they will tend to have similar joint distributions of observed covariates since PS is a balancing score. To check for the balance of the covariates, researchers suggest using standardized difference and graphic methods instead of inferential tests (Austin, 2009; Rosenbaum & Rubin, 1985; Rubin, 2001). Inferential tests are discouraged because they assume that there exists a superpopulation and they are affected by the sample size (Ho, Imai, King, & Stuart, 2007; Imai, King, & Stuart, 2008).

2.2.1 Propensity Score Estimation

2.2.1.1 Logistic Regression

PSs need to be estimated in observational studies given that the true PS is almost always unknown in observational studies in practice and thus must be estimated (Steiner & Cook, 2013). To date, the most frequently used method to estimate PSs is logistic regression (LR; Guo & Fraser,

2010). In LR, the treatment status is regressed on observed baseline covariates. Thus, the predicted probabilities of the group membership are the PSs for a given set of covariates. In constructing a logistic PS model, an iterative process is recommended whereby covariates are adaptively chosen and updated based on the improvements to the covariate balance (Rosenbaum & Rubin, 1984).

Despite its popularity, LR has limitations that make it less appealing for PS estimation (Westreich, Lessler, & Funk, 2010). First, LR requires the researchers to select covariates and specify the correct functional form for the PS model. However, when many covariates are available, it is difficult for researchers to identify which variables are important and include all the high-order terms to model nonlinear and/or nonadditive relationships between the treatment and the predictors. Yet incorrect PSs obtained from the misspecified PS model could result in a biased effect estimate (Drake, 1993). Second, LR cannot handle data with missing covariates or too few events per covariate.

2.2.1.2 Machine Learning Approaches

To overcome these limitations, non-parametric machine learning methods have been suggested as possible alternative approaches. In the context of PS estimation, machine learning approaches have several advantages over LR (Westreich et al., 2010). They are more flexible and require fewer assumptions than LR (Westreich et al., 2010). They can automatically deal with continuous and categorical covariates or any transformation of the covariates (such as log or square transformation). They can also automatically handle missing data by using surrogate predictors to classify cases with missing values on the predictors. Furthermore, they can automatically select variables for the PS model and capture higher-order relationships such as interaction and/or quadratic terms.

A number of machine learning algorithms have been applied to estimate PS, including classification and regression trees (CART) and ensemble methods, such as bagging, random forests, and generalized boosted models (GBM) (Westreich et al., 2010). Of these methods, GBM (Ridgeway, 1999; McCaffrey, Ridgeway, & Morral, 2004), a variant of boosting algorithm, was developed specifically for PS estimation.

2.2.1.3 LR Versus Machine Learning Approaches for PS Estimation

Some empirical and simulation studies have attempted to compare the machine learning methods with LR for PS estimation. Overall, bagging and (pruned) CART performed relatively poorly with respect to covariate balance, bias, and precision (Lee, Lessler, & Stuart, 2010; Luellen, 2007; Luellen, Shadish, & Clark, 2005; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Watkins et al., 2013). GBM produced mixed results. Some studies have shown that GBM performs well for PS estimation in certain settings compared with other machine learning methods and LR models, particularly when applied to weighting (e.g., Harder, Stuart, & Anthony, 2010; Lee et al., 2009; McCaffrey et al., 2004). In contrast, some reported that it exhibited poor performance when paired with matching (especially nearest neighbor matching) (Diamond & Sekhon, 2013; Pirracchio, Petersen, & van der Laan, 2015; Stone & Tang, 2013) and when used in a high-dimensional setting (Hill, Weiss, & Zhai, 2011). Random forests performed well in all the studies that examined it, including even in the high-dimensional settings (Cham, 2013; Keller, 2013). In conclusion, among the CART and ensemble methods, GBM and random forests stand out as promising alternatives to LR for PS estimation. In contrast, (pruned) CART and bagging are not recommended as PS estimation strategies.

2.2.2 Propensity Score Application

There are four techniques developed to use the estimated PSs to control for confounding (Schafer & Kang, 2008; Stuart, 2010). Of these four approaches, the most popular one in education and social science is matching, followed by stratification and weighting (Thoemmes & Kim, 2011; Hernandez, 2015). Note that any PS application method can be combined with any PS estimation method (Harder et al., 2010; Rosenbaum & Rubin, 1983).

2.2.2.1 Matching

PS matching refers to processes that match treated and control subjects based on their PSs. Two matching algorithms are generally used in practice. One is nearest neighbor (NN) matching algorithm, in which a treated subject is matched to an untreated subject whose PS is closest to that of the treated subject. The other is optimal matching algorithm, which intends to find matched pairs so as to minimize the total distance across all matched pairs. Previous studies have suggested that, for pair matching with a large group of controls, both algorithms create similar matched groups, but optimal matching results in better matched pairs (Gu & Rosenbaum, 1993). However, when there is a lack of control group members, NN matching can do much worse than optimal matching (Hansen, 2004). Rosenbaum (1989) has found that optimal matching is always at least as good as matching with greedy algorithm. A special type of optimal matching is full matching (Rosenbaum, 1991). In full matching, a treated unit can be matched with several control units, and several treated units can be matched with one control unit.

Treatment effects are estimated based on the matched groups. Note that the pair matching generally estimates ATT, and full matching can estimate both ATE and ATT depends on the choice of weights (Stuart, 2010). One major problem with pair matching is the loss of subjects (Schafer

& Kang, 2008), which would influence the causal estimand of the study and result in slight reduction in power.

2.2.2.2 Stratification

Stratification, also called subclassification, is a method that forms groups (subclasses) of all treated and control subjects with similar PSs (Rosenbaum & Rubin, 1984). The optimal number of strata depends on the sample size and the degree of overlap between the treatment and control groups' PSs (Stuart, 2010). In general, five strata are used because approximately 90% of the bias due to covariates can be removed (Cochran, 1968; Rosenbaum & Rubin, 1984). Stratification can estimate either ATE or ATT depending on the weights used in combining stratum estimates (Stuart, 2010). Compared to matching, stratification uses all subjects, but there is a risk of very unbalanced comparison groups in some strata when the sample size is relatively small.

2.2.2.3 Weighting

The idea of reweighting treated and control subjects by corresponding PSs to make them more representative of the population of interest was proposed by Rubin (2001). This idea is the same as for inverse-probability weighting in survey research (Horvitz & Thompson, 1952). Two major schemes have been developed for PS weighting. One is known as inverse probability of treatment weighting (IPTW) (Lunceford & Davidian, 2004), which weights treated and control subjects to represent the population. This weighting scheme is used to estimate ATE. Specifically, a treated unit receives a weight of the inverse of its estimated PS, and a control unit receives a weight equals to the inverse of 1-PS:

$$w_1 = \frac{1}{PS} \tag{2}$$

$$w_0 = \frac{1}{1-PS} \quad (3)$$

The other one is called weighting by the odds (Hirano, Imbens, & Ridder, 2003), which weights the control group to resemble the treatment group. This weighting scheme can estimate ATT. Specifically, a treated unit receives a weight of 1, and a control unit receives a weight equals to the inverse of 1-PS first and then multiplies its PS to mimic the treated group:

$$w_1 = 1 \quad (4)$$

$$w_0 = \frac{PS}{1-PS} \quad (5)$$

The derived weights can then be used in a weighted least squares regression model to estimate treatment effect. Like stratification, weighting retains all subjects. However, a potential drawback of the weighting method is the possibility of very large weights that results from extreme PSs (i.e., if the estimated PS is close to 0 or 1). If the model used to estimate PS is correct, then these large weights are accurately derived and, thus, pose no harm (Stuart, 2010). Otherwise, these large weights may reduce the precision of the estimates of treatment effects (Cole & Hernán, 2008, in Austin & Stuart, 2017).

2.2.2.4 Covariate Adjustment Using the Propensity Score

PSs can also be included as a covariate in the outcome model, either alone, or along with other covariates. This approach sometimes produced unbiased effect estimates (e.g., Austin, Grootendorst, & Anderson, 2007). However, many researchers argue against such use of propensity scores (Stuart, 2010; Schafer & Kang, 2008). First, this approach relies on the correct specification of PS and of the outcome model. Second, it does not take advantage of the balancing property of PS. Third, it is not clear which type of treatment effect is estimated using this method.

According to the author's literature review, using PSs as a regression covariate is rarely seen in recent educational empirical studies.

In summary, each method of constructing treatment and control groups using PSs has its own advantages and limitations. A PS adjustment technique that could produce uniformly superior performance in all settings does not exist. Moreover, there is a strong interaction between the PS estimation method and the application method on the effect estimation (e.g., Austin et al., 2007; Harder et al., 2010; Luellen, 2007). Thus, the choice of the specific estimation and adjustment technique in practice depends on the data as well as the type of treatment effect researchers intend to estimate.

2.2.3 PS Implementation in Practice

For many researchers, a big challenge in using observational data is whether they can use QE methods to obtain unbiased estimates of treatment effects. Starting with Lalonde (1986), researchers have attempted to use RCTs to validate QE designs. In these studies, researchers estimate the program's impact using randomized comparison groups from an RCT, then re-estimate the impact by using nonrandomized comparison groups and QE techniques. The results from the RCT provide a benchmark for evaluating whether the QE designs can recover the treatment effect. If the results are similar, it suggests that the adjustment to the QE is effective. This type of research is called the within-study comparison (WSC) study. The QE approaches that have been examined in the literature include regression-discontinuity, interrupted-time series, matching (including PS approaches), difference-in-difference, and standard regression adjustment. The present review focuses on the WSC studies in the education context that have applied propensity scores as one of the QE approaches in their WSC designs.

The evidence on the ability of PS adjustments to approximate experimental results is mixed. Some WSC papers report that PS methods perform well against an experimental benchmark (e.g., Bifulco, 2012; Shadish, Clark, & Steiner, 2008), whereas others fail to recover the experimental impacts (e.g., Agodini & Dynarski, 2004; Wilde & Hollister, 2007). Cook, Shadish, & Wong (2008) and Wong et al. (2017) provide excellent reviews of WSC studies in the fields of education, social science, and health. These two reviews, as well as some WSC and simulation[2] research, attempt to identify the factors that would influence bias reduction in field settings using QE methods. These factors are presented as follows.

2.2.3.1 Pretests

Pretests, or proxy pretests, often reduce bias in observational studies (Bifulco, 2012; Cook & Steiner, 2010). However, their actual performance depends on its correlation with the treatment selection and outcomes (Steiner, Cook, Shadish, & Clark, 2010). Steiner et al. (2010) reanalyzed the data from Shadish et al. (2008) and found that (proxy) pretests did not perform well if they had weak relationships with the treatment and outcomes. In contrast, for strong correlations between pretests and treatment and outcomes, even a single pretest measure can be enough to eliminate almost all of the bias in an observational study (Aiken, West, Schwalm, Carroll, & Hsiung, 1998; St. Clair, Cook, & Hallberg, 2014).

2.2.3.2 Rich Covariate Set

The strong ignorability assumption underlying PS implies that all covariates related to both outcomes and treatment assignment need to be observed and included for PS estimation

[2] The simulation studies here refer to those that built their simulations on the data used in WSC studies.

(Rosenbaum & Rubin, 1983). When researchers have full knowledge of selection process and have collected a rich set of covariates that are related to both treatment and outcomes, QE methods generally work well (e.g., Shadish et al., 2008; Pohl, Steiner, Eisermann, Soellner, & Cook, 2009). Here, the richness in a so-called rich set of covariates is defined by the combination of the number of construct domains as well as the number of covariates per domain (Steiner, Cook, Li, & Clark, 2015). The simulations by Steiner et al. (2015) suggest that the number of heterogeneous construct domains is more important for bias reduction than the number of covariates within each domain.

In reality, however, researchers generally do not tend to have a strong theory of selection process or access to a rich set of covariates. Oftentimes, only demographic variables are available, such as gender, age, ethnicity, disability status, and free-reduced lunch status. Research has shown that PS methods that only rely on demographic covariates produce biased results (Bifulco, 2012; Steiner et al., 2010; Wilde & Hollister, 2007).

2.2.3.3 Critical Covariates

Critical covariates refer to the set of covariates that is most predictive of treatment assignment and outcomes. Two WSC studies (Steiner et al., 2010; Hallberg, 2013) found that the use of the most critical covariate sets can be as effective as combining all covariates to reduce selection bias. In Steiner et al. (2010), each of the two critical constructs (one consisting of one covariate item and the other consisting of two covariate items) removed almost all the bias that all the 156 covariate items combined could remove. In Hallberg (2013), each of the two critical constructs (one consisting of two waves of math and reading pretests and the other consisting of two waves of teacher pretreatment evaluations of student performance) removed nearly as much bias as all the 208 covariates combined. Even though the datasets used in these two WSC studies contained at least 150 covariates, Steiner et al. (2010) used data from Shadish et al (2008), in which

researchers had a strong selection theory and collected all the covariates on their own, whereas Hallberg (2013) had to rely on the covariate information a national dataset could provide. One simulation study (Steiner, Cook, Li, & Clark, 2015) based on these two datasets supported the findings from Steiner et al. (2010) and Hallberg (2013). This finding is very important, as it suggests that, when researchers have no idea which covariates are critical, they can use all the available covariates to estimate PSs as long as the sample size does not prevent them from developing complex PS models.

2.2.3.4 Reliability of Covariates

The reliability in the covariates also affects the performance of PS adjustment to reduce bias (Cook & Steiner, 2010; Steiner, Cook, & Shadish, 2011). If a covariate is a true confounder (i.e., related to both treatment and outcomes), its potential to reduce selection bias increases as its reliability increases. In contrast, if a covariate is not related to either treatment or outcomes, the measurement error in this covariate does not affect bias reduction. However, it may reduce the precision of the effect estimates (Steiner & Cook, 2013). In sum, reliability in the covariates is important for bias reduction, but not as important as the choice of the right covariates (Cook & Steiner, 2010; Steiner et al., 2011).

2.2.3.5 Local Matching

Local matching suggests matching treatment and comparison groups within the same geographic area. There are a broad range of definitions of the word “local” in the WSC studies. Local may refer to the same school (Fortson, Verbitsky-Savitz, Kopa, & Gleason, 2012), school district (Bifulco, 2012), or even state (Dong & Lipsey, 2014). According to Cook et al. (2008),

local matching, in theory, not only reduces bias from observed covariates but also from the unobserved covariates that are related to treatment and outcome.

Although local PS matching is advocated in the literature, its actual performance depends on whether comparable local comparison groups can be found. If not, local matching may generate undesirable results. For example, Bifulco (2012) found that the non-local comparisons with similar observed characteristics performed better than the local comparisons with different observed characteristics. Dong and Lipsey (2014) had similar findings. When a treated unit cannot find comparable local matches, a *hybrid* matching strategy is recommended (Stuart & Rubin, 2008). In hybrid matching, treated units attempt to first find local matches that have similar important observed characteristics. For those without comparable local matches, non-local matches will be identified based on important observed characteristics.

In sum, the review of WSC studies in the education context has shown that the success of observational methods such as PS adjustment depends on whether the researcher has identified all important covariates related to treatment selection and outcomes and has measured them reliably. Among all the covariates, pretests need special attention. If possible, local matching needs to be considered prior to non-local matching. Once these factors are satisfied, the choice of analytic method does not really matter for bias reduction (e.g., Cook & Steiner, 2010; Shadish et al., 2008; Steiner et al., 2010; Steiner et al., 2011).

2.3 PROGNOSTIC SCORE ADJUSTMENT

An alternative to PS analysis is based on prognostic scores (PROGs). A PROG is formally defined as any scalar or multidimensional function that, when conditioned on, induces

independence between measured covariates and the potential outcomes under control conditions (Hansen, 2008):

$$Y_c \perp X | \Psi(X) \quad (6)$$

where Y_c is the potential outcomes under control conditions, X refers to the set of observed covariates, and $\Psi(X)$ is the PROG. The PROG is called a "disease risk score" (DRS) if the outcome is binary (Arbogast & Ray, 2009; Glynn, Gagne, & Schneeweiss, 2012). The PROG generalizes the DRS to continuous, categorical, and ordinal outcomes. As a "prognostic analogue" of the PS (Hansen, 2008), the PROG is also a data-reduction technique that combines multiple covariates into a single score[3]. However, rather than modeling the association between treatment and covariates as in PSs, PROGs model the relationship of covariates and potential outcomes in the absence of treatment (i.e., suppose each subject is in the control condition).

The PROG is estimated by fitting a prognostic model (i.e., the outcome model) in the control group and then using this fitted model to estimate expected outcomes (i.e., PROGs) under control conditions for all treatment and control members. The estimated PROGs are then used to adjust for selection bias, using methods similar to those used with PSs (Hansen, 2008).

Similar to PS, a well-formed PROG also has a prognostic balance (Hansen, 2008). The balancing property of PS means that on average, units with similar PSs have similar distributions of covariates that contributed to PSs (Rosenbaum & Rubin, 1983). Thus, if matching or stratification on PSs could be achieved, the covariates would have the same distributions across the treatment conditions, thus eliminating bias in the treatment effect estimates due to differences

[3] Unlike PS, PROG theoretically can be a scalar or a vector-valued function of the confounders (Hansen, 2008). Specifically, if the outcome is binary, the PROG is a scalar. If the outcome is continuous, the PROG can be a scalar or a vector-valued function of the confounders. However, according to my literature research, PROGs are all used as a scalar in either empirical investigations or methodological discussions. Therefore, PROGs are viewed as a scalar in my dissertation.

in the covariates. In parallel, prognostic balance suggests that conditional on PROGs, the potential outcome under the control conditions would be independent of the covariates included in the prognostic model (Hansen, 2008). Balance on PROGs would indicate balance on the covariates highly predictive of the outcome (Hansen, 2008). If the balance on these covariates could be achieved through stratifying or matching on PROGs, the bias and variance of the treatment effect estimates would be reduced because the covariates highly predictive of the outcome could cause the most bias if left unbalanced (Hansen, 2008). This is a major benefit of adjusting for PROGs. However, unlike PS balance, the prognostic balance can be checked only in the control condition. One problem is that the procedures for checking for the prognostic balance have not been fully developed.

There is also an inherent problem in estimating PROGs with the sample used to fit prognostic models. In theory, the PROG models could be fitted using either full sample or control sample only (Hansen, 2008). Even though full sample estimation has been shown to outperform control-only estimation in some conditions (Arbogast & Ray, 2011; Cadarette et al., 2010), Hansen (2008) argued for using the control sample instead of the full sample. He explained that fitting the prognostic model using the full sample would make the fitted values dependent on treatment status. Therefore, the effect estimates would be biased. This bias would worsen if the covariates differed substantially between the treatment groups.

Fitting the RPOG model only in the control group would also give rise to another problem – overfitting (Hansen, 2008). Overfitting suggests that the model fit would be inherently better for the control group than for the treatment group. Such overfitting could cause spurious prognostic score differences between the treated and the control subjects, and it could potentially bias the overall effect estimates (Glynn et al., 2012; Hansen, 2008). In the literature, PROGs are assumed

to be the control-only prognostic scores, and they are prone to overfitting. At present, settings in which the overfitting issue would be a problem is not clear.

Regardless of using the full sample or the control sample only, the same sample is used for both PROG and the treatment effect estimation. Hansen (2008) referred to this inherent problem with the estimation of PROGs as the "same-sample" estimation problem. To avoid this problem, Hansen (2008) and Glynn et al. (2012) suggested using a separate sample or historical data from the study population to fit the prognostic model. They then suggested using the study population to estimate treatment effects. In practice, it is difficult to find a separate or historical sample that has similar characteristics of the study population.

2.4 COMBINING PROPENSITY SCORES AND PROGNOSTIC SCORES

Hansen (2008) viewed PROG as a complement rather than an alternative to PS. In the DRS literature, the estimated DRS is generally used as a continuous or a categorical (e.g., quintiles, deciles) covariate in the outcome model to control for confounders (Arbogast & Ray, 2009; Glynn et al., 2012). Hansen (2008) argued that such use of PROGs did not take advantage of PROGs' balancing property as PROGs are mainly used as a dimension reduction tool. Instead, he suggested matching or stratifying on PROGs. Furthermore, Hansen (2008) proposed the use of PROGs in combination with PSs due to the inherent estimation issues with PROGs. Hansen suggested that, theoretically, conditioning on both scores would yield more precise effect estimates than conditioning on PSs alone did when there was a limited overlap of PSs between treatment and control groups.

A limited number of studies have examined the joint use of PSs and PROGs in the estimation of treatment effects. To the best of my knowledge, only four methodological studies (Hansen, 2006; Kelcey & Swoboda, 2015; Leacy & Stuart, 2014; Tu & Koh, 2017) have investigated the methods combining propensity and prognostic scores in observational studies. Of these four studies, two are simulations (Leacy & Stuart, 2014; Tu & Koh, 2017), one is a case study (Kelcey & Swoboda, 2015), and one conducts both (Hansen, 2006).

Hansen (2006) proposed the first method of combining the two scores: full matching on a Mahalanobis distance combining the estimated PSs and PROGs. In other words, the estimated PSs and PROGs were combined into a Mahalanobis distance and then full matching on this distance was implemented to match the treatment and control subjects. In a case study evaluating the effect of the SAT coaching program, Hansen found that his newly proposed two-score method resulted in similar estimates to those from a PS analysis, but with smaller standard errors (SEs). In the simulation study, he considered performing the full matching first with no caliper restrictions and then with propensity or prognostic score calipers. Consistent with the results from the empirical example, the simulations showed that, with calipers or not, the combination of the two scores through a Mahalanobis distance yielded less biased and more efficient effect estimates than full matching on PSs only.

Hansen (2008) also proposed another method of combining PSs and PROGs to reduce selection bias. This approach used the estimated PROGs to estimate PSs with PROGs as the only predictor in the PS model, or as a predictor along with other observed covariates in the PS model. The predicted PS through this method is called the prognostic propensity score (ProgPS) by Hansen (2008). This method was proposed as a means to improve propensity balance on prognostically important covariates. Tu and Koh (2017) investigated the performance of this

method relative to adjustments on PSs or PROGs alone. Specifically, Tu and Koh (2017) conducted Monte Carlo simulations to compare four different summary scores for adjusting bias in estimating the marginal and conditional rate ratios of count data. The four estimated summary scores were PSs, PROGs, and two ProgPSs (the estimated PROGs as the sole predictor, or the estimated PROGs and other covariates together as the predictors). A total of 11 matching adjustment strategies were applied to each summary score.

Method performance was evaluated in terms of bias and mean squared error (MSE). Results showed that the key factor of reducing bias was the choice of adjustment algorithm. Matching with replacement was preferred for estimating the marginal rate ratio, whereas matching without replacement was preferred for estimating the conditional rate ratio. Paired with the right algorithm, all four summary scores performed similarly in terms of bias and MSE when other factors were held constant. This seemingly lack of advantage for ProgPSs over single score adjustment might be due to the simulation settings used in the study. The simulations in Tu and Koh (2017) only considered true confounders and correctly specified PS and PROG models. Under such settings, adjustments using ProgPSs may not provide any additional bias reduction relative to single score adjustment. The two-score methods might outperform single score methods in other conditions, such as in the presence of PS and/or PROG model misspecification.

Leacy and Stuart (2014) conducted a series of simulations to evaluate the robustness of three methods of combining PSs and PROGs in the presence of the misspecification of propensity and/or prognostic models. Of these three methods, two were the methods proposed by Hansen: full matching on a Mahalanobis distance combining the estimated PSs and PROGs, and full matching on the estimated ProgPSs within PS calipers. Leacy and Stuart (2014) proposed their own approach of combining propensity and prognostic scores: subclassification on an estimated propensity and

prognostic score grid with 5 x 5 subclasses. This method divides the data in a 5x5 grid of subclasses based on the quintiles of the estimated propensity and prognostic scores. Similar to stratification on PSs or PROGs alone, the treatment effect is calculated as a weighted average of the within-subclass estimates.

Leacy and Stuart (2014) compared these three two-score methods with a number of single score methods and with standard main effects linear regression analysis. They examined the performance of each method in four scenarios: with either the PS or the PROG model being misspecified and with both models being either correctly specified or misspecified. Moreover, the true PS and PROG models varied in the degrees of linearity and/or additivity. Performance across the methods was evaluated in term of bias, SE, root mean squared error (RMSE), and 95% confidence interval coverage rates.

Simulation results showed that the joint use of PS and PROGs, especially the two methods proposed by Hansen, exhibited excellent and robust performances across all simulation settings and scenarios. When both PS and PROG models were misspecified, Hansen's two methods were found to outperform all single score adjustments. When both score models were correctly specified, they were only marginally outperformed by full matching on PROGs. When only one of the score models was correctly specified, their performance was almost as effective as the single score adjustment. Specifically, in scenarios where the PS model was correctly specified while the PROG model was incorrectly specified, Hansen's methods were almost as effective as full matching on PSs for effect estimation. In scenarios where the PS model was incorrectly specified while the PROG model was correctly specified, they were almost as effective as full matching on PROGs for estimating treatment effects. The stratification method of combing two scores the authors proposed did not perform as well as Hansen's two-score methods. As to the relative

performance of Hansen's two methods, both showed comparable performances in most cases, but full matching on a Mahalanobis distance combining the estimated PSs and PROGs was slightly better in the scenarios when only the PROG model was correctly specified.

To sum up, the findings from Leacy and Stuart (2014) suggest that the joint use of PSs and PROGs may be preferred to single score adjustment for effect estimation in observational studies. Full matching on PROGs alone did exhibit the best performance when the PROG model was correctly specified. However, Hansen's two methods may still be preferred in practice because it is very difficult for researchers to be absolutely certain about their prognostic models. Hansen's two methods showed strong performance in settings where only one of the score models or neither models were correctly specified. In other words, combing PSs and PROGs could protect against the misspecification either of the PS model or the PROG model or both.

Kelcey and Swoboda (2015) extended the joint use of PSs and PROGs to multilevel settings. They proposed an alternative means of combining two summary scores—PS matching within PROG strata. This approach refers to stratifying on PROGs (generally five strata) and then performing 1:1 NN matching on PSs with replacement within PROG strata. Kelcey and Swoboda (2015) illustrated the use of their method in a multilevel setting using data from Early Childhood Longitudinal Study-Kindergarten (ECLS-K). For simplicity, the authors only considered four pretreatment demographic covariates. They compared the performance of this method with PS matching alone. Given multilevel data, multilevel PROGs were estimated using random intercept and slope (RIS) model. Then, the entire sample was subclassified into five groups based on students' multilevel PROGs. Students were then matched based on their PSs within a caliper of 0.1. A single-level logistic PS model that included interactions and polynomials was used. In this illustrative sample, both matching on PS alone and PS matching within PROG strata were

conducted across clusters instead of within clusters. Treatment effect was estimated with a multilevel RIS model. With respect to the estimation of the treatment effects, both approaches produced different effect estimates but the same SEs. Because the true treatment effect was unknown, which adjustment approach produced better effect estimates was unclear. However, the authors demonstrated that their proposed technique of combining both scores could improve the balance on each covariate across treatment conditions, much like adjustment on PSs alone. The improved balance across treatment conditions suggests that the overt bias (i.e., the bias caused by the observed covariates) may be reduced. Therefore, the authors argued that PS matching within PROG strata could be a promising strategy to reduce bias and variance of effect estimates even in multilevel settings.

In summary, four major methods of using PROGs in combination with PSs have been proposed in the literature: Mahalanobis distance combining the estimated PSs and PROGs (Hansen, 2006), ProgPSs (Hansen, 2008), PS matching within prognostic strata (Kelcey and Swoboda, 2015), and stratifying on both PSs and PROGs (Leacy & Stuart, 2014). Overall, these approaches, especially Hansen's two methods, have shown strong potential for treatment effect estimation. They exhibited better performance than single score adjustment especially in the presence of model misspecification. Full matching on PROGs alone might be a better choice if researchers are absolutely confident in the correctness of their PROG models. However, given the robustness found with Hansen's two methods (Leacy & Stuart, 2014), applying Hansen's methods for effect estimation is better than applying single score adjustment.

3.0 METHODOLOGY

A simulation study was conducted to compare the separate and combined use of propensity and prognostic scores for treatment effect estimation in quasi-experimental (QE) designs. The current simulation study was based on restricted-use data from an impact evaluation study of the U.S. Department of Education's Student Mentoring Program (SMP) (Bernstein, Rappaport, Olsho, Hunt, & Levin, 2009). This dataset was chosen for several reasons. First, it provided a realistic educational context. While simulation methods based on artificial data could also provide insights about methods, how well those simulation settings approximate reality is unclear. Moreover, such an approach does not address many issues that researchers encounter in practical evaluation settings. Second, the SMP evaluation study was a multisite randomized controlled trail (RCT), with sites "purposively selected" and students at each site selected into the study. Thus, the SMP data could be directly used to construct a QE design. Specifically, a specific site was selected as a treatment site, and all the students from the selected site served as the treatment group in the QE design. The comparison group population in the QE design consisted of the students from the remaining sites. Given that treatment and control students came from different sites, they were unlikely to affect each other. Thus, the stable unit treatment value assumption for propensity score (PS) adjustment was very likely satisfied because the chance that the treatment students from one site would interact with their potential comparisons from other sites was minimal (Stuart, 2010). Third, the SMP data provided a relatively rich set of covariates, which implies a higher chance of meeting the strong ignorability assumption for PS adjustment.

3.1 DATA SOURCE FOR SIMULATION

The SMP was a federally funded program that provided school-based mentors to at-risk students in grades 4-9. Starting from 2005, a large-scale multisite RCT evaluation was conducted to evaluate the effectiveness of the school-based student mentoring programs on a range of student outcomes (Bernstein, et al., 2009). Of the 255 SMP grantees (i.e., sites) that received federal funding in 2004 or 2005, only 32 met the research team's selection criteria and were purposively selected for the evaluation. Two cohorts of students were sampled for the study. The first cohort of students was recruited from 21 sites in Fall 2005. The second cohort was recruited from 21 sites in Fall 2006. Of these 42 grantees, 10 provided students in both years, but students who participated in the first study cohort were excluded from the second cohort. Therefore, 42 distinctive groups of students from 32 unique sites were included in the sample. Students were randomly assigned to treatment and control groups at each site. In total, the 32 grantees recruited 2,573 students who were randomly assigned either to a treatment group (n = 1,272) or to a control group (n = 1,301).

The SMP data came from four sources:

- Student school records (Cohort 1: Fall 2005 and Spring 2006; Cohort 2: Fall 2006 and Spring 2007)
- Student survey (Cohort 1: Fall 2005 and Spring 2006; Cohort 2: Fall 2006 and Spring 2007)
- Mentor survey (Cohort 1: Spring 2006; Cohort 2: Spring 2007)
- Grantee survey (Cohort 1: Spring 2006; Cohort 2: Spring 2007)

Out of three surveys, mentor and grantee surveys were collected posttreatment. Only the student surveys were collected both pretreatment (mostly prior to treatment assignment) and posttreatment. The following student-level background variables were collected:

- Gender
- Age
- Race/Ethnicity
- Free or reduced lunch status (FRL)
- Family structure (whether coming from two-parent households)
- Prior mentoring experience (whether having mentors the previous school year)
- Prior mentoring experience frequency
- Whether receiving mentoring at least two times a month
- Had academic risk
- Had disciplinary risk

In addition, the study measured 17 outcomes in three crucial domains:

- Interpersonal relationships and personal responsibility
- Academic achievement and engagement
- High-risk or delinquent behavior

These 17 outcome measures were derived from both student surveys and school records. The SMP study estimated 42 site-specific impacts and an overall impact across sites for each of the 17 outcomes. The researchers did not find any statistically significant overall[4] impact on any of the 17 outcomes after controlling for multiple comparisons.

[4] Bernstein et al. (2009) did not report site-specific impacts for each outcome. Only the overall impact was reported.

3.2 SIMULATION DESIGN

3.2.1 Design Factors

Two criteria guided the choice of design factors: (1) investigate factors that have not been studied, or factors that have been found to affect the effect estimation of the adjustment methods in the previous within-study comparison studies and PS literature, and (2) mirror the conditions that are found in applied educational research. Based on these criteria, four factors were examined in the current study.

3.2.1.1 Propensity Score Estimation Methods

Two methods were compared: (1) logistic regression (LR), and (2) generalized boosted models (GBM). Logit models were performed in SAS 9.4 (SAS institute, 2017). The GBM algorithm was implemented using the R package *twang* (Ridgeway, McCaffrey, Morral, Burgette, & Griffin, 2017). For *twang*, recommended settings (i.e., the number of regression interactions = 3 and shrinkage value applied at each iteration of the algorithm = 0.005) were used (Ridgeway et al., 2017). A simulation study by Austin (2012) has shown that these settings work well.

LR was selected because it is the most commonly used method for estimating PSs. GBM was selected for two reasons. First, GBM is the first (and probably the only) machine learning algorithm to be specifically developed for PS estimation. Second, GBM is a promising alterative to LR for estimating PS according to prior studies (e.g., Harder et al., 2010; Lee et al., 2009; McCaffrey et al., 2004).

3.2.1.2 Methods for Constructing Comparison Groups

In the current study, average treatment effect on the treated (ATT) was estimated using the sample of the treatment group at each site and the comparison groups constructed using different samples and methods. The following eight methods of constructing comparison groups were examined:

(1) 1:1 optimal pair matching on the estimated prognostic scores (PROGs);

(2) 1:1 optimal pair matching on the estimated linear PSs;

(3) ATT weighting on the estimated PSs;

(4) 1:1 optimal pair matching on the estimated linear prognostic propensity scores (ProgPSs);

(5) ATT weighting on the estimated ProgPSs;

(6) 1:1 optimal pair matching on a Mahalanobis distance combining the estimated linear PSs and PROGs;

(7) Full matching on a Mahalanobis distance combining the estimated linear PSs and PROGs;

(8) Naïve method.

Methods (1) – (7) were adjustment methods using either PSs, PROGs, or both PSs and PROGs. Method (8) was included for comparison purposes. It did not adjust for any pretreatment covariate differences. Instead, the raw differences of outcomes between treatment and control groups were calculated. Optimal matching was performed using R package *Optmatch* through *MatchIt* (Ho, Imai, King, & Stuart, 2007). All other methods were implemented in SAS 9.4 (SAS institute, 2017).

There are many different matching algorithms and matching strategies available for use. Pair matching was selected because matching (either nearest neighbor (NN) or optimal) is a dominant PS approach in applied educational research, particularly the pair matching (Hernandez, 2015). Moreover, compared to NN matching, optimal algorithm performs better when the sample size of controls is limited (Gu & Rosenbaum, 1993; Hansen, 2004). Therefore, this study is focused on optimal pair matching even though NN may be more popular than optimal matching among applied researchers (Stuart, 2010). As for the methods that combine PSs and PROGs, Leacy and Stuart (2014) and Hansen (2006) found that one of the two-score methods (i.e., full matching on a Mahalanobis distance combining the estimated PSs and PROGs) showed strong performance when using the settings evaluated in their studies. Thus, method (7) was evaluated in this study to explore its potential in the educational settings. In addition, given that the current study focused on optimal pair matching, optimal pair matching on the same distance measure was also examined (method (6)).

Several decisions were made with respect to the specific matching implementations. First, no caliper was imposed on the matching in the current study even though a caliper is often applied in PS matching in practice (Austin, 2009; Stuart, 2010). This study attempted to estimate the ATT on *all* treated students. However, the use of a caliper could result in the loss of treated students due to a lack of common support. This would change the estimand from the average effect on *all* treated students to the effect on a *subgroup* of treated students with certain PSs. In contrast, removing the caliper restriction maintained the same estimand across different adjustment methods. Furthermore, even when matching within the caliper restriction shows good bias reduction, there is not much value when the inferences can be generalized only to a very restricted sample. Therefore, all the matching in this study was implemented without caliper.

Second, the logit scale of PS (i.e., $\hat{l}(X) = log\{PS/(1 - PS)\}$), also called linear PS, was used when matching on PSs was involved (methods (2), (4), (6), and (7)). Two reasons justified the choice of the linear PS. First, compared to PS itself, the linear PS is recommended for PS matching or covariance adjustment because it is more linearly related to the outcome and it is also more normally distributed (Rosenbaum & Rubin, 1985; Rubin, 2001; Schafer & Kang, 2008; Steiner & Cook, 2013). Second, prior studies have shown that Mahalanobis distance matching performs well when there are relatively few covariates (less than five) and when these covariates are approximately normally distributed (Gu & Rosenbaum, 1993, Rubin & Thomas, 2000). Thus, this study transformed the estimated PS to its logit scale for matching. Note that all matching was performed without replacement because matching with replacement is only applicable to NN matching algorithm (Bai, 2015; Stuart & Rubin, 2008).

Weighting was chosen because no researchers have yet investigated its use with ProgPSs (method (5)). Moreover, weighting maintains all the sample size, and weighting is easier to implement than matching or stratification. Thus, its performance relative to other approaches is of great interest to applied researchers. For comparison purposes, weighting on PSs (method (3)) was also performed. Given the interest of the current study was on the ATT estimates, ATT rather than ATE weighting scheme was used.

Stratification is also one popular PS adjustment method. However, stratification applications were not investigated in the current study for two reasons. First, the sample size available may not allow for its implementation (Kelcey & Swoboda, 2015). Second, stratifying on either PROGs only or both PSs and PROGs was found to be inferior to their respective full matching counterparts (Leacy & Stuart, 2014).

3.2.1.3 Control Group Size

As mentioned, comparison groups were constructed using students from sites other than the sites from which the treated students were selected. For example, if the treated students from one site (Site 1) were selected as the treatment group for the QE design, then the control students from other sites (Sites 2 – *N*) formed the potential pool of comparison group members. The PS was then estimated using the combined sample from the treated students in Site 1 and all the control students from Sites 2 - *N*.

In the current study, the sample size of control groups included three levels: (1) 3-site without contextual factor condition, (2) 3-site with contextual factor condition, and (3) 38-site condition. For the 38-site control group condition, all members from the unselected sites were used as the pool of comparison group members. For the 3-site without contextual factor condition, three sites were randomly selected from the unselected sites (i.e., not including the site selected for treatment group) and all students from these three sites were used to form the population of comparison group members. For the 3-site with contextual factor condition, three sites were randomly selected from the unselected sites that shared the same category of contextual factor as the treatment site, and then all students from these three sites were used to form the population of comparison group members.

The 3-site and 38-site conditions were selected for two reasons. First, it allowed for examining the influence of the sample size ratios between the treatment and control group on the effect estimation of adjustment methods. According to my literature review, treatment exposure level has not been commonly reported in the empirical studies published in educational journals. Of those that reported the level, ratios of treated to control units were usually around 1:1 ~ 1:3[5].

[5] Simulation studies on PSs generally set treated-to-control ratios at either 1:3 or 1:4 (e.g., Leacy & Stuart, 2014).

Thus, the 3-site condition was selected to approximately reflect the ratio range found in empirical educational research. The 38-site condition was used to create a scenario in which control group size was much larger than treatment group size. Note that regardless of control group size, the current study formed a research context consisting of small educational evaluations due to the relatively small sample size of treated students at each site.

Second, it allowed for exploring how two-score methods and PROG adjustment performed with a small sample size. Many educational evaluations use small sample sizes. However, PS analysis is essentially a method for large samples (Rubin, 1997). Previous simulation studies suggest that sample sizes larger than $N = 1000$ are desirable, or at least with $N = 500$ (Luellen, 2007; Lee et al., 2010). According to Hernandez (2015), none of the QE studies published in the four top education journals[6] from 2012 to 2014 had an overall sample size smaller than 600. All prior simulation studies on methods combining both PSs and PROGs had sample sizes of $N \geq 1000$. However, the joint use of both scores or PROG adjustment may be vulnerable to small sample sizes just like PS applications. For either of the 3-site conditions in this study, the total sample size was less than 400. Therefore, the 3-site condition may address the total sample size question. In summary, the 3-site and 38-site control group conditions attempted to explore the influence of both total sample size and sample size ratio on the performance of different adjustment methods.

In addition, a contextual factor was introduced into the 3-site condition because it was hypothesized that the students' observed and unobserved background characteristics between the treatment and control groups might be more similar to each other if the students came from the

[6] These journals are *American Educational Research Journal*, *Educational Evaluation and Policy Analysis*, *Research in Higher Education*, and *Journal of Research on Educational Effectiveness*.

sites sharing the same category of the contextual factor. As a result, more accurate effect estimates may be obtained due to the higher degree of similarity between treatment and control group members. This hypothesis can be tested by comparing the results from the two 3-site conditions. In this study, grantee's experience of running school-based mentoring programs was selected as the contextual factor. The original variable was a continuous variable indicating the number of years in running the program, ranging from less than one year to 30 years with a mean of six years. For the study purpose, I categorized this continuous variable into quartiles.

3.2.1.4 Type of Outcome Domains

One within-study comparison study by Griffen and Todd (2017) revealed that there may be larger variation in bias across outcome domains than variation in bias across adjustment methods. They found that regardless of the method used, larger biases were observed for income and employment outcomes than for child test scores and child health outcomes. Thus, this study sought to explore whether the performance of methods differed depending on the outcome types.

In the current study, outcomes from two domains (one academic achievement outcome and one delinquent behaviors outcome) were chosen. These two domains are also the dominant outcome domains in applied educational research. Several criteria were considered for the selection of specific outcome measures. First, one school-reported outcome measure and one self-reported outcome measure were selected. Second, the preference was for outcome measures that had the least missing data on their pretest measures. Third, this study was restricted to continuous outcomes.

Given these criteria, the self-reported academic measure "school efficacy and bonding" and school-reported delinquent behaviors outcome "absenteeism rate" were chosen as the two outcome measures. The "school efficacy and bonding" measure was a composite score derived

from eight 4-point Likert scale items. It had an acceptable reliability with a Cronbach's alpha of 0.72 (Bernstein et al., 2009). It also had a low missingness rate of 3.8% at baseline. Moreover, its distribution was less skewed than another academic self-reported measure "future orientation". The outcome "absenteeism rate" was chosen because it had the lowest missingness rate (20.2%) among all school-reported delinquent behaviors outcomes.

In sum, the design factors—size of control groups (three levels) and outcome domains (two levels)—resulted in six conditions. Each of these six conditions was analyzed with 14 nonexperimental methods, for a total of 84 combinatory analyses of treatment recovery. The 14 nonexperimental methods consisted of 13 adjustment methods and one naïve method without any adjustment of covariates. The 13 adjustment methods were formed by crossing two estimation methods with six methods of constructing comparison groups plus one additional adjustment using PROGs only. Of these 13 adjustment methods, four methods used only PSs, one used only PROGs, and eight used both PSs and PROGs. A list of the 14 nonexperimental methods is provided in Table 1.

Table 1. Nonexperimental Methods Applied to Simulated Data

Nonexperimental Methods
1. 1:1 optimal pair matching on PROGs
2. ATT weighting on the LR-estimated PSs
3. ATT weighting on the LR-estimated ProgPSs
4. 1:1 optimal pair matching on the LR-estimated linear PSs
5. 1:1 optimal pair matching on the LR-estimated linear ProgPSs
6. 1:1 optimal pair matching on a Mahalanobis distance combining the LR-estimated linear PSs and PROGs
7. Full matching on a Mahalanobis distance combining the LR-estimated linear PSs and PROGs
8. ATT weighting on the GBM-estimated PSs
9. ATT weighting on the GBM-estimated ProgPSs
10. 1:1 optimal pair matching on the GBM-estimated linear PSs
11. 1:1 optimal pair matching on the GBM-estimated linear ProgPSs
12. 1:1 optimal pair matching on a Mahalanobis distance combining the GBM-estimated linear PSs and PROGs

Table 1 continued

13. Full matching on a Mahalanobis distance combining the GBM-estimated linear PSs and PROGs
14. Naïve method

Note. PROG = Prognostic score; ATT = Average treatment effect on the treated; LR = Logistic regression; PS = Propensity score; ProgPS = Prognostic propensity score; GBM = Generalized boosted models.

3.2.2 Preprocessing Data

In order to use the SMP data as a basis for simulation, the empirical example data were preprocessed using the following procedures.

3.2.2.1 Missing data

Some covariates in the SMP data contained missing values. Of all available student covariate data, the only covariates that did not contain missing data were “age”, “gender”, and “had disciplinary risk”. Even though GBM can handle missing data automatically, LR needs complete data. Therefore, missing covariate values were imputed. In the current study, the missing data imputation procedures matched those of several methodological PS investigations (Cham, 2013; Hallberg, 2013; Steiner et al., 2015). In their respective studies, researchers used a singly imputed data set from multiple imputation. They based their decision on the small number of missing values and a purely methodological interest in analyzing the different methods’ abilities to remove bias. Since missing data existed for both continuous and categorical covariates, I imputed the missing values using multiple imputation methods via fully conditional specification method in SAS 9.4 (SAS institute, 2017) with the number of imputations set to one. The fully conditional specification method is also known as the chained equations method in R. This

algorithm was chosen because it can address continuous, binary, ordered categorical, and count data (Cham, 2013; Cham & West, 2016). The list of imputed variables was as follows[7]:

- Race/Ethnicity (African American, Hispanic, White, and American Multirace/Other)
- FRL (Yes = 1, No = 0)
- Two-parent households (Yes = 1, No = 0)
- Prior mentoring experience (Yes = 1, No = 0)
- Had academic risk (Yes = 1, No = 0)
- Grades (math, reading/ELA, science, and social studies)
- Pro-social behaviors
- Future orientation
- Student efficacy and bonding
- Absenteeism rate

Table 2 and Table 3 show the characteristics of these imputed covariates before and after imputation, respectively. I used standardized bias (SB) to assess the differences between the treatment and control students on these covariates (Austin, 2011). As shown in Table 2, all these student pretreatment covariates had baseline equivalence, as the absolute values of their SBs were all smaller than 0.10. After I imputed the missing data for these covariates, I retested their balance and found that the distributions between the treatment and control groups were again similar in the imputed sample, with no absolute SBs greater than 0.10 (see Table 3). Furthermore, the means,

[7] Note that two student-level background covariates were not selected for score estimation due to their extremely high missingness rate: "prior mentoring experience frequency" (75%) and "whether receiving mentoring at least two times a month" (75%). In addition, several baseline measures of outcomes were not selected either because of their high missingness rates: "truancy rate" (44%), and school-reported misconduct and delinquency measures (28%). Thus, these covariates were not imputed.

standard deviations, and SBs for each continuous covariate, as well as the percentages and SBs for each categorical covariate, were all very similar before and after imputation. This suggested that the imputation of missing covariates was successful. The imputed sample had a sample size of 2,573 students as the original sample.

Table 2. Descriptive Statistics of Student-Level Imputed Covariates before Imputation

	Treat		Control		Standardized Bias (SB)	% of Missing
	Categorical Covariates					
	N	Percentage	*N*	Percentage		
Race/Ethnicity						
African American	1207	42	1254	40	0.04	4.4
Hispanic	1207	29	1254	33	-0.09	4.4
White	1207	24	1254	21	0.07	4.4
American Multirace/Other	1207	5	1254	6	-0.02	4.4
FRL (Yes = 1)	1125	84	1151	88	-0.11	11.5
Two-parent households (Yes = 1)	1254	56	1291	57	-0.01	1.1
Prior mentoring experience (Yes = 1)	1240	27	1280	26	0.02	2.1
Had academic risk (Yes = 1)	977	60	1034	60	0.00	21.8
	Continuous Covariates					
	N	Mean	*SD*	*N*	Mean	*SD*	Standardized Bias (SB)	% of Missing
Grades								
Math	850	3.31	1.02	916	3.29	1.07	0.02	31.4
Reading/ELA	865	3.48	1.01	917	3.44	1.02	0.03	30.7
Science	849	3.51	1.06	908	3.49	1.00	0.02	31.7
Social studies	837	3.47	1.05	882	3.50	1.03	-0.02	33.2
Pro-social behaviors	1228	2.87	0.54	1258	2.86	0.52	0.01	3.4
Future orientation	1233	3.84	0.38	1262	3.81	0.46	0.07	3.0
School efficacy and bonding	1216	3.15	0.55	1258	3.15	0.54	0.00	3.8
Absenteeism rate	1023	0.05	0.06	1029	0.05	0.06	-0.02	20.2

Note. The formula for calculating standardized bias for continuous variables is $SB = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{S_T^2 + S_C^2}{2}}}$ (Austin, 2011), where $\bar{X}_T$ and $\bar{X}_C$ denote the mean of the variable in the treatment and control groups, respectively, and S_T^2 and S_C^2 denote the variance of the variable in the treatment and control groups, respectively. The formula for calculating standardized bias for categorical variables is $SB = \frac{\hat{P}_T - \hat{P}_C}{\sqrt{\frac{\hat{P}_T(1-\hat{P}_T) + \hat{P}_C(1-\hat{P}_C)}{2}}}$ (Austin, 2011), where $\hat{P}_T$ and $\hat{P}_C$ denote the prevalence or mean of the dichotomous variable in the treatment and control groups, respectively.

Table 3. Descriptive Statistics of Student-Level Imputed Covariates after Imputation

	Treat		Control		Standardized Bias (SB)
	Categorical Covariates				
	N	Percentage	*N*	Percentage	
Race/Ethnicity					
African American	1272	42	1301	40	0.05
Hispanic	1272	29	1301	33	-0.10
White	1272	24	1301	21	0.06
American Multirace/Other	1272	5	1301	6	-0.03
FRL (Yes = 1)	1272	84	1301	87	-0.09
Two-parent households (Yes = 1)	1272	57	1301	57	-0.00
Prior mentoring experience (Yes = 1)	1272	27	1301	26	0.02
Had academic risk (Yes = 1)	1272	60	1301	60	-0.00

	Continuous Covariates						Standardized Bias (SB)
	N	Mean	*SD*	*N*	Mean	*SD*	
Grades							
Math	1272	3.32	1.04	1301	3.30	1.06	0.02
Reading/ELA	1272	3.46	1.02	1301	3.45	1.04	0.02
Science	1272	3.51	1.05	1301	3.49	1.02	0.02
Social studies	1272	3.51	1.05	1301	3.48	1.07	0.03
Pro-social behaviors	1272	2.87	0.53	1301	2.87	0.52	0.00
Future orientation	1272	3.84	0.39	1301	3.81	0.45	0.08
School efficacy and bonding	1272	3.16	0.56	1301	3.15	0.54	0.01
Absenteeism rate	1272	0.05	0.06	1301	0.06	0.06	-0.03

Note. The formula for calculating standardized bias for continuous variables is $SB = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{S_T^2 + S_C^2}{2}}}$ (Austin, 2011), where $\bar{X}_T$ and $\bar{X}_C$ denote the mean of the variable in the treatment and control groups, respectively, and S_T^2 and S_C^2 denote the variance of the variable in the treatment and control groups, respectively. The formula for calculating standardized bias for categorical variables is $SB = \frac{\hat{P}_T - \hat{P}_C}{\sqrt{\frac{\hat{P}_T(1-\hat{P}_T) + \hat{P}_C(1-\hat{P}_C)}{2}}}$ (Austin, 2011), where $\hat{P}_T$ and $\hat{P}_C$ denote the prevalence or mean of the dichotomous variable in the treatment and control groups, respectively.

3.2.2.2 The Covariates

In the present simulation study, 11 student-level pretreatment covariates from the SMP data were used for PS/PROG estimation:

- Age
- Female (Yes = 1, No = 0)
- Two-parent households (Yes = 1, No = 0)
- Prior mentoring experience (Yes = 1, No = 0)
- Had academic risk (Yes = 1, No = 0)
- Had disciplinary risk (Yes = 1, No = 0)
- GPA
- Pro-social behaviors
- Future orientation
- Student efficacy and bonding
- Absenteeism rate

Of these 11 covariates, GPA was the only covariate that was not created by the SMP study investigators. I created the proxy GPA variable by averaging the imputed values of students' baseline grades on math, reading/ELA, science, and social studies. Variable "female" was derived from the dichotomous variable "gender". The selected outcome measures were "student efficacy and bonding" and "absenteeism rate". Therefore, their baseline measures served as the pretests. The study investigators created the variable "had academic risk" based on students' performance on two state assessment variables and the variable "had disciplinary risk" based on self-reported outcome measures "misconduct" and "delinquency". Other available student level covariates like "prior mentoring experience frequency", "whether receiving mentoring at least two times a

month", "truancy rate" and school-reported misconduct and delinquency were not selected because of their high missingness rates at baseline (44% - 75%). After excluding the variables with high missingness rates, the variables based on which selected covariates were created, and the pretest measures of selected outcomes, two self-reported outcome measures remained: "pro-social behaviors" and "future orientation". Both variables were composite scale scores and had low missingness rates. The former was derived a ten 4-point Likert scale items and the latter was derived from three 4-point Likert scale items. Both had acceptable internal consistencies (Cronbach alpha was 0.70 and 0.76, respectively). Thus, the pretest measures of these two variables were selected for estimating PSs or PROGs.

Of these 11 covariates, five were binary variables, including "female", "two-parent households", "prior mentoring experience", "had academic risk", and "had disciplinary risk". All binary covariates were dummy coded 0 and 1. The remaining covariates were continuous. To reduce the impact of scale differences across the variables, all continuous covariates were standardized.

It should be noted that some site-level covariates were also available in the original SMP data. I compared two hierarchal linear models using the simulated outcomes: one hierarchical linear model without any covariates and one with all selected 11 student-level covariates. I found that student-level covariates explained almost all of the site level variance (~ 99%) in most of the constructed QE data. Therefore, only student-level covariates were included in the PS/PROG estimation models.

3.2.2.3 Analytic Sample Used in the Study

The preliminary multivariable LR analyses in approximately one third of the replications did not converge because there was little variability in the FRL and minority[8] (i.e., non-White) variables: in these non-converged replications, 98 - 99% of the students were minority students (with the exception of one replication, which had 89% minority students), and 92-100% of students were FRL students (with the exception of four replications, which had 62- 85% FRL students). Because 71% minority students received FRL in the imputed full sample[9], restricting the sample to minority students who participated in the FRL program would not significantly change the main characteristics of the sample for these two variables and would not substantially reduce the sample size. Thus, I restricted the imputed sample to minority students participating in the FRL program to ensure that students were matched explicitly on these two variables. As a result, the sample size was reduced from 2,573 to 1,830, with the sample size per site ranging from eight to 103. In addition, since the sample size within each site determined the treatment group size and small treatment size caused unstable estimates in preliminary multivariable LR analyses, the site-level sample size was also restricted. Based on the preliminary analyses, sites with at least 11 students were used in the study. Thus, three sites were removed, which further reduced the total sample size from 1,830 to 1,802. The N = 1,802 from 39 sites formed the final analytic sample for the simulation study.

[8] The variable "minority" was created from the imputed Race/Ethnicity variable: minority = 1 for non-White, and minority = 0 for White.
[9] This imputed sample had 78% minority students and 85% FRL students.

3.2.3 Simulation Procedure

3.2.3.1 Outcome Simulation

In this study, a treatment effect to the treated students was simulated according to the following outcome generating model:

$$Y_i = Pretest_i + d * Z_i + \varepsilon_i \quad (7)$$

where Y_i is the simulated continuous outcome for student *i*, Z_i is an indicator of treatment status (1 = treatment; 0 = control) for student *i*, d is the average treatment effect of 0.30, and $\varepsilon_i \sim N(0, \sigma^2)$. The error term σ is given by $\sigma = \sqrt{\frac{n_t+n_c}{n_t n_c} + \frac{d^2}{2(n_t+n_c)}}$ (Hedges & Olkin, 1985) where n_t and n_c are the sample sizes for the treatment and control groups, respectively. In the current study, the simulated effect size d = 0.30 served as the true effect benchmark against which all nonexperimental methods were compared.

3.2.3.2 PS/PROG/ProgPS Estimation

The same 11 covariates were used to estimate either PSs or PROGs so that the use of covariates was not a confounder. The PS estimation model is given in Equation 8:

$$\ln(\frac{e(x)}{1-e(x)}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{11} X_{11} \quad (8)$$

where $e(x)$ is the estimated propensity score, and X_1 to X_{11} refer to the 11 selected modeling covariates. Note that Model 8 specifically refers to logistic estimation of PSs. For GBM, same 11 covariates were used to estimate PSs but no functional form needed to be specified. The PROG estimation is based on the following outcome model:

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_{11} X_{11} \quad (9)$$

where X_1 to X_{11} refer to exactly the same 11 covariates in Model 8, and Y_i refers to each student's observed outcome under the control condition. After the regression coefficients β were determined by fitting Model 9 to the control group students, I used the fitted model to estimate the prognostic scores for all the treatment and control group students. To estimate ProgPSs, due to the convergence issues[10], the model includes only two covariates: the estimated PROG and the pretest of the outcome measures:

$$\ln(\frac{e(x)}{1-e(x)}) = \beta_0 + \beta_1 * X_1 + \beta_2 * PROG \quad (10)$$

where $e(x)$ is the estimated ProgPS, X_1 refers to the pretest of the outcome measure, and PROG is the estimated PROG based on Model 9. Again, Model 10 refers to the logistic estimation only. The pretest measure was included because it was considered as the critical covariate for PS estimation and was highly recommended to be included if it was available (e.g., Steiner et al., 2010).

Note that the logistic PS, the logistic ProgPS, and the PROG models included only main effect terms of the selected covariates. The choice of main effect terms was justified by two reasons. First, it is very difficult to construct different logistic PS or ProgPS models using the recommended iterative process in a simulation study. Second, building a main-effects only LR model is a usual practice for applied researchers when they conduct a PS analysis (Thoemmes & Kim, 2011). In other words, including only main effects for each of the covariates could reflect

[10] First, the estimated PROGs and the 11 covariates used to estimate PROGs were used together to estimate ProgPSs. However, this estimation did not converge because of the relatively large number of covariates and the relatively small treatment group size in some replications. Thus, the final ProgPS estimation model consisted of only two covariates: the estimated PROG and the pretest of the outcome measures. Given that the PROG estimation process already contained all 11 covariates, it was acceptable to use only the pretest and the estimated PROGs to estimate ProgPSs.

practical situations in which the true functional relationships between covariates and treatment were unknown.

3.2.3.3 Treatment Effect Estimation

Once each matched/weighted comparison group was constructed, a standardized mean difference (i.e., Cohen's *d*) in outcomes between treatment and comparison group was calculated as the estimated treatment effect. In this simulation study, the treatment was not intended to impact the variance of the simulated outcomes (Model 7). Therefore, using Cohen's *d* as a measure of treatment effect was reasonable. To better isolate the effects of adjustment methods, no covariates were controlled after adjustment was applied (e.g., Lee et al., 2010).

For each constructed QE, effect size *d* was calculated as follows:

$$d = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{(n_T-1)SD_T^2+(n_C-1)SD_C^2}{n_T+n_C-2}}} \quad (11)$$

where $\bar{X}_T$ and $\bar{X}_C$ are values of outcome means for students in the treatment and control groups, respectively, SD_T^2 and SD_C^2 are values of the outcome variance for students in the treatment and control groups, respectively, and n_T and n_C are the number of students in the treatment and control groups, respectively. Note that Hedges' formula for pooled standard deviation was used to calculate effect size in Equation 11. For full matching and weighting adjustments, weighted standardized effect size was calculated in SAS 9.4 (SAS institute, 2017). Specifically, the weighted mean is given by $\bar{X}_{weight} = \frac{\sum w_i x_i}{\sum w_i}$, and the weighted variance is given by $SD_{weight}^2 = \frac{1}{n-1}\sum w_i \, (x_i - \bar{X}_{weight})^2$ (SAS institute, 2017). The w_i is the weight assigned to each student from the full matching and weighting procedures. For full matching, the weight for each treatment subject is 1 and the weight for each control student is proportional to the number of treated subjects

divided by number of control subjects within each matched set (Stuart, 2010). For weighting, each treatment subject receives a weight of 1 and each control subject receives a weight of $\frac{PS}{1-PS}$ (Stuart, 2010).

3.2.4 Data Generation and Analysis

Steps for data generation and analysis included:

(1) Select one outcome measure.

(2) Start with Site 1 and select all students within Site 1 as the treatment group (note: all students from the remaining 38 sites are treated as control students).

(3) Simulate a random small standardized effect size (mean d = 0.30) as the treatment effect on the students in the treatment group (note: the outcomes of both treated and control students (mean d = 0.00) are simulated in this step).

(4) Select all students from the remaining 38 sites to form the pool of potential comparison group members, OR, randomly select 3 sites from 38 sites and use all students from these 3 selected sites to form the pool of potential comparison group members, OR, randomly select 3 sites from the sites that share the same category of contextual factor as the selected treatment site, and then use all students from these 3 selected sites to form the pool of potential comparison group members.

(5) For the selected treatment group members and potential comparison group members from Step (4), estimate PROGs/PSs/ProgPSs for each treated and control student using the selected covariates. For PS estimation, use each of the two estimation methods (LR versus GBM).

(6) Using the PS/PROG/ProgPS estimates from Step (5), create matched/weighted comparison groups based on each of the proposed adjustment methods.

(7) Based on the matched/weighted samples from Step (6), estimate a (weighted) standardized effect size and assess covariate balance using the selected balance check criterion. The selected balance check criterion was the absolute standardized bias in the covariates (Stuart, 2010).

(8) Repeat Steps (2) - (7) to obtain the estimates of treatment effect and covariate balance statistics when sites 2, 3 … 39 were selected as the treatment site sequentially. Recovery of the simulated treatment effect was evaluated by examining the bias and root mean squared deviation of the simulated treatment effect across the 39 replications. The covariate balance was evaluated by the average standardized bias for each of the 11 covariates and across all 11 covariates across the 39 replications.

(9) Repeat Steps (1) – (8) for another selected outcome measure.

3.2.5 Study Outcome Measures

The main purpose of this study was to compare the abilities of different adjustment methods in recovering the simulated treatment effect. The recovery of the simulated treatment effect for each of ATT estimator was evaluated by examining the bias and root mean squared deviation (RMSD) of the effect size estimates across 39 replications. Due to the varying sample sizes for each constructed QE, a standard meta-analytic technique (Borenstein, Hedges, Higgins, & Rothstein, 2010; Lipsey & Wilson, 2001) was used to calculate the bias and RMSD of effect sizes in each condition. Specifically, the bias and RMSD across replications were calculated by:

$$Bias = \frac{\sum_{r=1}^{39}(w*(\hat{d}-.3))}{\sum_{r=1}^{39} w} \quad (12)$$

$$RMSD = \sqrt{\frac{\sum_{r=1}^{39}(w*(\hat{d}-.3)\text{^}2)}{\sum_{r=1}^{39} w}} \quad (13)$$

where $\hat{d}$ is the estimated effect size, and w is a fixed effect inverse variance weight. The fixed effect inverse variance weight was calculated as:

$$w = \frac{1}{\sigma^2} \quad (14)$$

where σ^2 is the within study variance of the effect size estimate. The within study variance formula is as follows:

$$\sigma^2 = \frac{n_T+n_C}{n_T n_C} + \frac{\hat{d}^2}{2(n_T+n_C)} \quad (15)$$

where $\hat{d}$ is the estimated effect size, n_T is the treatment group sample size, and n_C is the control group sample size.

In addition to the bias and RMSD, this study examined the balance in the covariates between treated and control students to explore the relationship between the balance in covariates' distributions and the bias/RMSD of the ATT estimates. In this study, the balance for individual covariates and across covariates in each condition was assessed in the original, matched, and weighted samples using the commonly recommended balance measure absolute standardized bias (ASB):

$$ASB = \frac{|\bar{X}_T-\bar{X}_C|}{SD_T^{un}} \quad (16)$$

where $\bar{X}_T$ is the treatment group sample mean of covariate *X*, $\bar{X}_C$ is the control group sample mean of covariate *X*, and SD_T^{un} is the unadjusted standard deviation of covariate *X* in the treatment group. A lower ASB indicates better balance: the treatment and comparison groups are more similar with respect to the given covariate. There are no strict cut-off ASB values to indicate imbalance. Popular

cutoff values include 0.05 (Caliendo & Kopeinig, 2008), 0.10 (Normand et al., 2001), and 0.25 (Ho et al., 2007; Stuart & Rubin, 2008). As What Works Clearinghouse's guideline (2017) adopts the 0.25 threshold to indicate baseline equivalence (the treatment and control groups are considered to be equivalent if the ASBs of the baseline covariates are 0.25 or smaller), this study also used 0.25 as a cutoff value: an ASB smaller than 0.25 indicated adequate balance.

Note that Equation 16 is recommended specifically for checking covariate balance when ATT is the estimand of the study (Harder et al., 2010; Stuart, 2010; West et al., 2014). This formula is also adopted to provide the balance statistics in the *MatchIt* package (Ho et al., 2011) and in the *twang* package when the estimand is ATT (Ridgeway et al., 2017). For full matching and weighting methods, weights are applied to calculate the sample mean estimates. The weighted mean is defined as $\bar{X}_{weight} = \frac{\sum w_i x_i}{\sum w_i}$ (SAS institute, 2017), where w_i is the weight assigned to each student from the full matching and weighting procedures. Consistent with the calculation of the bias and RMSD of effect estimates, the ASB across replications in a simulation condition was also calculated using the standard meta-analytic technique:

$$ASB = \frac{\sum_{r=1}^{39}(w*\widehat{ASB})}{\sum_{r=1}^{39} w} \quad (17)$$

where $\widehat{ASB}$ is the estimated ASB for each covariate or across covariates in each replication, $w = \frac{1}{\sigma^2}$, and $\sigma^2 = \frac{n_1+n_2}{n_1 n_2} + \frac{\hat{d}^2}{2(n_1+n_2)}$.

4.0 RESULTS

In this chapter, I present the simulation results in two major sections: recovery of the simulated treatment effect size and the covariate balance between the treatment and control groups. Note that no model convergence problems were evident with the results discussed herein. A replication converged when the models for estimating the propensity scores (PSs), prognostic scores (PROGs), or prognostic propensity scores (ProgPSs) in SAS or *twang* converged, and when no error messages occurred when implementing matching in *MatchIt*. Based on these criteria, every replication in the simulation study converged. Thus, the results were all based on 39 replications.

4.1 RECOVERY OF THE SIMULATED TREATMENT EFFECT SIZE

The effect size recovery was evaluated based on the bias and the root mean squared deviation (RMSD) of the effect size estimates. The calculated raw bias measures the average tendency for the estimated effect (below or above the true effect) to assess whether an adjustment method would over- or underestimate the treatment effect. The RMSD is a combination of bias and variance and is a measure of the overall variability of the effect estimates. Examining RMSD could help us to identify the methods that produce the most precise estimates of the treatment effects.

4.1.1 Bias

Table 4 includes bias results for the naïve method (with no covariate adjustment) and the values from the application of the different propensity/prognostic scoring procedures. These results are presented for each combination of the PS estimation methods (logistic regression and generalized boosted models), outcomes measures, and control group size conditions.

4.1.1.1 Academic Outcome

As shown in the top panel of Table 4, for the academic outcome, the naïve method and all matching adjustments using PSs, PROGs, or both scores tended to yield a similar bias to one another across all control group size and PS estimation method conditions, with bias ranging from -0.07 to 0.08. Of these methods, the two-score methods based on a Mahalanobis distance—1:1 or full optimal matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG)—tended to slightly overestimate the effect sizes. Other matching methods tended to slightly underestimate the effect sizes under the two 3-site conditions and to slightly overestimate the effects under the 38-site condition.

Compared to the matching adjustments, weighting adjustments—weighting on the estimated PSs (W.PS) and weighting on the estimated prognostic propensity scores (W.ProgPS)—produced a much larger bias regardless of the PS estimation methods, particularly under the 38-site condition. The bias for W.PS and W.ProgPS ranged from 0.12 to 0.22 under the 3-site without contextual factor condition, from 0.15 to 0.28 under the 3-site with contextual factor condition, and from 1.07 to 1.29 under the 38-site condition. These large positive biases suggested that W.PS

and W.ProgPS showed a tendency to strongly overestimate the simulated effect sizes for the academic outcome, particularly under the 38-site condition.

The results also showed that for the academic outcome, the performance of W.PS and W.ProgPS was sensitive to the PS estimation methods. For both W.PS and W.ProgPS, logistic regression (LR) performed slightly better than the generalized boosted models (GBM) as weighting on the LR-estimated PSs or ProgPSs resulted in a slightly smaller bias than weighting on the GBM-estimated scores across all control group size conditions.

4.1.1.2 Disciplinary Outcome

As shown in the bottom panel of Table 4, for the disciplinary outcome, all adjustment methods appeared to overestimate the treatment effects, as each method had positive biases in all cases, with the bias ranging from 0.03 to 0.20. Although the estimates for each method were more biased for the disciplinary outcome than for the academic outcome, a similar pattern was found for both outcomes regarding the performance of the weighting adjustments relative to other methods. Compared to other methods, W.PS and W.ProgPS yielded larger biases across all control group sizes and PS estimation methods, particularly under the 38-site condition (for which the bias was also greater than 1 in all cases).

The results also revealed the influence of the PS estimation methods on the performance of the adjustment methods. For some adjustment methods, the PS estimation methods' impact was consistent across the control group size conditions, as it was for the academic outcome. For instance, across all control group size conditions, W.PS and W.ProgPS consistently yielded less biased effect sizes when paired with LR than when paired with GBM. For other adjustment methods, the influence of the PS estimation methods differed by the levels of the control group size factor. For instance, for 1:1 optimal matching on the estimated prognostic propensity scores

(1:1M.ProgPS), the LR-estimated ProgPS produced much more biased estimates than the GBM-estimated ProgPS did under the two 3-site conditions. On the other hand, the LR-estimated ProgPSs produced much lower bias than the GBM-estimated ProgPSs did under the 38-site condition. Similarly, for 1:1 optimal matching on the estimated propensity scores (1:1M.PS), both LR and GBM produced similar bias under the two 3-site conditions, but GBM produced larger bias than LR did under the 38-site condition.

Another result worth noting was that, regardless of the PS estimation methods, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG produced slightly smaller bias under the 3-site with contextual factor condition (bias = 0.03) than under the 38-site condition (bias = 0.07 ~ 0.08), and both biases were much smaller than the bias under the under the 3-site without contextual factor condition (bias = 0.15 ~ 0.19). This result might suggest that the selected contextual factor was useful for 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG to reduce bias for the disciplinary outcome.

Finally, it should be noted that the bias results for the weighting methods differed considerably in magnitude from the results for other methods, particularly for the 38-site control group condition. This was due to the way in which the pooled standard deviation in the effect size was calculated. To account for the differences in sample sizes between the treatment and control groups, the pooled standard deviation was calculated using the following formula (Hedges, 1981, p.110):

$$SD_{pooled} = \sqrt{\frac{(n_T-1)SD_T^2+(n_C-1)SD_C^2}{n_T+n_C-2}} \quad (18)$$

However, the pooled standard deviation may also be calculated by a simple average of the variances (Cohen, 1988, p.44):

$$SD_{pooled} = \sqrt{\frac{SD_T^2 + SD_C^2}{2}} \quad (19)$$

Bias in recovery values using Cohen's formula are provided in parentheses in Table 4. As can be seen these values are reduced remarkably. Regardless of which formula was used, inferences about the method comparison remain unchanged. Note that when the sample sizes between the groups are the same, Hedges' formula reduces to Cohen's formula.

Table 4. Bias of the Estimated Effect Sizes Across 39 Replications

Size of Control Groups (Number of Sites)	Naïve Method	1:1M.PROG	LR						GBM					
			W.PS[a]	W.ProgPS[a]	1:1M.PS	1:1M.ProgPS	1:1MAHAL .PS.PROG	FULL.MAHAL. PS.PROG	W.PS[a]	W.ProgPS[a]	1:1M.PS	1:1M.ProgPS	1:1MAHAL. PS.PROG	FULL.MAHAL. PS.PROG
Academic Outcome														
3 (without contextual factor)	-0.02	-0.02	0.12 (0.06)	0.14 (0.08)	-0.07	-0.02	-0.02	0.00	0.22 (0.08)	0.18 (0.10)	-0.03	-0.03	0.03	0.05
3 (with contextual factor)	0.02	0.01	0.15 (0.07)	0.16 (0.09)	0.00	-0.02	0.03	0.03	0.28 (0.11)	0.22 (0.12)	-0.02	-0.04	0.06	0.07
38	0.00	0.03	1.07 (0.15)	1.07 (0.15)	0.01	0.02	0.07	0.08	1.29 (0.12)	1.20 (0.14)	0.05	0.00	0.07	0.08
Disciplinary Outcome														
3 (without contextual factor)	0.09	0.14	0.16 (0.09)	0.18 (0.12)	0.09	0.12	0.15	0.16	0.35 (0.19)	0.23 (0.14)	0.10	0.04	0.16	0.19
3 (with contextual factor)	0.06	0.13	0.14 (0.06)	0.15 (0.07)	0.11	0.14	0.03	0.03	0.38 (0.18)	0.23 (0.13)	0.09	0.03	0.03	0.03
38	0.00	0.09	1.08 (0.15)	1.10 (0.18)	0.07	0.08	0.06	0.08	1.75 (0.25)	1.45 (0.21)	0.13	0.20	0.07	0.10

Note. LR = Logistic regression; GBM = Generalized boosted models; 1:1M.PROG = 1:1 optimal matching on prognostic scores; W.PS = Weighting on propensity scores; W.ProgPS = Weighting on prognostic propensity scores; 1:1M.PS = 1:1 optimal matching on propensity scores; 1:1M.ProgPS = 1:1 optimal matching on prognostic propensity scores; 1:1MAHAL.PS.PROG = 1:1 optimal matching on a Mahalanobis distance combining propensity and prognostic scores; FULL.MAHAL.PS.PROG = Full matching on a Mahalanobis distance combining propensity and prognostic scores.

[a] Bias results without parentheses were based on effect sizes (d) estimated using formula $d = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{(n_T-1)SD_T^2+(n_C-1)SD_C^2}{n_T+n_C-2}}}$, where $\bar{X}_T$ and $\bar{X}_C$ denote the values of outcome means for students in the treatment and control groups, respectively, SD_T^2 and SD_C^2 denote the values of the outcome variance for students in the treatment and control groups, respectively, and n_T and n_C are the number of students in the treatment and control groups, respectively. The values in parentheses were based on effect sizes estimated using formula $d = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{SD_T^2+SD_C^2}{2}}}$.

4.1.2 RMSD

Table 5 includes RMSD results for the naïve method (with no covariate adjustment) and the values from the application of the different propensity/prognostic scoring procedures. These results are presented for each combination of the PS estimation methods (LR and GBM), outcomes measures, and control group size conditions.

4.1.2.1 Academic Outcome

As shown in the top panel of Table 5, of all methods, 1:1M.PROG tended to result in the effect size estimates with the lowest RMSDs, with the RMSD being 0.06 under the 38-site condition, 0.11 under the 3-site with contextual factor condition, and 0.19 under the 3-site without contextual factor condition. The RMSDs for 1:1M.PROG were much smaller than the RMSDs found for the naïve method, which was 0.29 under the 38-site condition, 0.32 under the 3-site with contextual factor condition, and 0.35 under the 3-site without contextual factor condition. Two other methods also consistently performed better than the naïve method: 1:1M.PS and 1:1M.ProgPS. These two methods produced comparable RMSDs that were smaller than those for the naïve method in all cases.

In contrast, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG consistently performed worse than the naïve method. For each level of the control group size conditions, the RMSDs for 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG were larger than those for the naïve method, regardless of the PS estimation methods.

The performance of W.PS and W.ProgPS depended on the control group size conditions. Under the 38-site condition, out of all methods, W.PS and W.ProgPS resulted in the estimates with the greatest RMSD, which were greater 1 across all PS estimation methods. Under the 3-site

conditions, compared to the naïve method, W.ProgPS resulted in smaller RMSD. W.ProgPS also had similar RMSDs to 1:1M.PS and 1:1M.ProgPS, regardless of the PS estimation methods. In contrast, W.PS performed no better than the naïve method under the two 3-site conditions; they had similar RMSDs when the PSs were estimated from LR, and W.PS had greater RMSDs than the naïve method when the PSs were estimated from GBM. Note that, under the two 3-site conditions, the RMSDs for W.PS was larger than those for 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG when PS was estimated from GBM.

In addition, for all matching methods, the RMSDs under the 38-site condition were smaller than those under the two 3-site conditions. However, the differences among control group size conditions for each method were modest at most. We also found that, for all methods that involved PSs or ProgPS, GBM produced slightly greater RMSDs than LR did across the control group size conditions.

4.1.2.2 Disciplinary Outcome

As shown in Table 5, the RMSD values for each method in each crossed condition were similar across the disciplinary and academic outcomes. Furthermore, the relative performance of methods for the disciplinary outcome was similar to that for the academic outcome:

(1) Among all methods, 1:1M.PROG tended to result in the lowest RMSD across the control group size and the PS estimation method conditions;

(2) Regardless of the PS estimation methods, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG resulted in RMSDs that were no lower than those for the naïve method across the control group size conditions;

(3) 1:1M.PS and 1:1M.ProgPS resulted in lower RMSDs than the naïve method across all conditions except for one: GBM-estimated ProgPSs had higher RMSD than the naïve method under the 38-site condition (0.41 versus 0.30);

(4) W.PS and W.ProgPS produced the largest RMSDs of all methods under the 38-site condition, with the RMSDs ranging from 1.36 to 1.93 across the PS estimation methods;

(5) Under the two 3-site conditions, W.ProgPS produced slightly lower RMSDs than the naive method across the PS estimation methods, but W.PS resulted in greater RMSDs than the naïve method, particularly when W.PS used the GBM-estimated PSs;

(6) For all methods involving PSs or ProgPSs, GBM tended to produce slightly larger RMSDs than LR across all the control group size conditions.

As for the academic outcome, the RMSDs for the disciplinary outcome were smaller under the 38-site condition than under the two 3-site conditions. The only exception occurred for 1:1M.ProgPS when ProgPS was estimated from GBM. For this method, the RMSDs under the 38-site condition was larger than those under the two 3-site conditions (0.41 vs. 0.25 (3-site without contextual factor) and 0.30 (3-site with contextual factor)). As for the two 3-site conditions, for the academic outcome, the RMSDs for all matching adjustments tended to be larger under the 3-site without contextual factor condition than under the 3-site with contextual factor condition, and vice versa for the disciplinary outcome. For all weighting adjustments, the RMSDs were largest under the 38-site condition, followed by the 3-site with contextual factor condition, and then the 3-site without contextual factor condition regardless of the outcomes.

Again, as with the bias results, the RMSDs for the weighting methods had smaller magnitudes when using Cohen's formula (as opposed to Hedges' formula) to calculate the pooled

standard deviation. This was particularly true for the 38-site control group condition. The RMSDs using Cohen's formula are provided in parentheses in Table 5. However, unlike the bias results, the two formulas' RMSD results provide different inferences for the method comparison. For example, in the 38-site control group condition, with Hedges' formula, the weighting methods were the worst performing of all the examined methods in terms of RMSDs. In contrast, with Cohen's formula, the weighting methods performed no worse than 1:1MAHAL.PS.PROG or FULL.MAHAL.PS.PROG. This result indicates that it is important to select the appropriate formula when calculating the pooled variance for weighting estimators. Otherwise, the inferences can differ.

4.1.3 Summary

In summary, all of the examined adjustment methods, particularly W.PS and W.ProgPS, showed a tendency to overestimate the simulated effect sizes across all conditions. Of all methods, 1:1M.PROG had the best performance in terms of effect size recovery, followed by 1:1M.PS and 1:1M.ProgPS. These three methods all outperformed the naïve method. In contrast, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG performed similarly poorly. Across all conditions, they performed no better than the naïve method. W.PS and W.ProgPS also performed poorly, particularly under the 38-site condition, for which they produced both extremely large biases and RMSDs; of these two methods, W.ProgPS was slightly better than W.PS.

I also evaluated the impact of several factors—the outcome measures, the PS estimation methods, and the control group size—on the performance of these adjustment methods. It was found that all methods tended to produce slightly less biased estimates for the academic outcome and slightly more efficient estimates using the LR-estimated PSs or ProgPSs. In addition, all

matching methods tended to produce slightly more efficient estimates under the 38-site condition than under the two 3-site conditions, whereas the estimates based on the weighting methods tended to be more efficient under the two 3-site conditions. Within the 3-site condition, the impact of the selected contextual factor on the efficiency of the effect size estimates was mixed for matching adjustments. For the academic outcome, slightly more efficient estimates were obtained with the contextual factor, whereas for the disciplinary outcome, slightly more efficient estimates were obtained without the contextual factor. In contrast, for weighting adjustments, the impact of the contextual factor was consistent across the outcome measures: slightly more efficient estimates were always obtained without the contextual factor. Note that for the disciplinary outcome, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG tended to yield much less biased but more inefficient estimates with the contextual factor.

Table 5. RMSD of the Estimated Effect Sizes Across 39 Replications

Size of Control Groups (Number of Sites)	Naïve Method	1:1M.PROG	LR						GBM					
			W.PS[a]	W.ProgPS[a]	1:1M.PS	1:1M.ProgPS	1:1MAHAL.PS.PROG	FULL.MAHAL.PS.PROG	W.PS[a]	W.ProgPS[a]	1:1M.PS	1:1M.ProgPS	1:1MAHAL.PS.PROG	FULL.MAHAL.PS.PROG
Academic Outcome														
3 (without contextual factor)	0.35	0.19	0.27 (0.23)	0.19 (0.12)	0.25	0.22	0.43	0.43	0.46 (0.29)	0.23 (0.13)	0.29	0.25	0.44	0.44
3 (with contextual factor)	0.32	0.11	0.35 (0.26)	0.22 (0.13)	0.18	0.15	0.37	0.37	0.53 (0.29)	0.27 (0.15)	0.25	0.18	0.39	0.41
38	0.29	0.06	1.12 (0.17)	1.12 (0.16)	0.11	0.12	0.31	0.32	1.49 (0.19)	1.31 (0.17)	0.16	0.21	0.32	0.33
Disciplinary Outcome														
3 (without contextual factor)	0.31	0.18	0.36 (0.28)	0.32 (0.25)	0.22	0.21	0.37	0.39	0.56 (0.34)	0.32 (0.22)	0.25	0.25	0.38	0.42
3 (with contextual factor)	0.37	0.22	0.41 (0.35)	0.36 (0.29)	0.26	0.24	0.53	0.56	0.55 (0.30)	0.34 (0.22)	0.29	0.30	0.51	0.53
38	0.30	0.15	1.36 (0.34)	1.36 (0.29)	0.19	0.17	0.29	0.33	1.93 (0.29)	1.71 (0.27)	0.27	0.41	0.29	0.34

Note. LR = Logistic regression; GBM = Generalized boosted models; 1:1M.PROG = 1:1 optimal matching on prognostic scores; W.PS = Weighting on propensity scores; W.ProgPS = Weighting on prognostic propensity scores; 1:1M.PS = 1:1 optimal matching on propensity scores; 1:1M.ProgPS = 1:1 optimal matching on prognostic propensity scores; 1:1MAHAL.PS.PROG = 1:1 optimal matching on a Mahalanobis distance combining propensity and prognostic scores; FULL.MAHAL.PS.PROG = Full matching on a Mahalanobis distance combining propensity and prognostic scores.

[a] RMSD results without parentheses were based on effect sizes (d) estimated using formula $d = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{(n_T-1)SD_T^2+(n_C-1)SD_C^2}{n_T+n_C-2}}}$, where $\bar{X}_T$ and $\bar{X}_C$ denote the values of outcome means for students in the treatment and control groups, respectively, SD_T^2 and SD_C^2 denote the values of the outcome variance for students in the treatment and control groups, respectively, and n_T and n_C are the number of students in the treatment and control groups, respectively. The values in parentheses were based on effect sizes estimated using formula $d = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{SD_T^2+SD_C^2}{2}}}$.

4.2 COVARIATE BALANCE

Tables 6 through 19 report the mean absolute standardized bias (ASB) between the treatment and control groups across 39 replications for each covariate, as well as the averaged ASB across covariates for each crossed condition of control group size and outcome measure. Recall that an ASB less than 0.25 indicates adequate balance (Ho et al., 2007; Stuart & Rubin, 2008). ASBs less than but close to 0.25 may still imply substantial selection bias, as the 0.25 threshold was an arbitrary choice (Harder et al., 2010). Of the 11 measured covariates, the focus was on the balance of the pretest measures of outcome. The pretest is important because it is the most prognostically important covariate (i.e., the one that is most strongly related to the outcome) for the study's generated outcomes (Model 7). In this simulation study, "student efficacy and bonding" is the pretest measure of the academic outcome, and "absenteeism rate" is the pretest measure of the disciplinary outcome.

4.2.1 Absolute Standardized Bias

4.2.1.1 Pre-Adjustment Balance

Table 6 shows the initial covariate balance before adjustment. Based on the unadjusted ASB values for each covariate, the observed characteristics were not well balanced between the treatment and control groups. Specifically, 10 covariates (91%) under the 3-site without contextual factor condition, 9 covariates (82%) under the 3-site with contextual factor condition, and 7 covariates (64%) under the 38-site condition had ASBs greater than 0.25. As shown in Table 6, no

covariates had unadjusted ASBs lower than 0.22. The covariate with the largest imbalance was "age"[11] (ASB = 1.03 under the 3-site without the contextual factor condition; ASB = 1.11 under the 3-site with the contextual factor condition; ASB = 0.88 for the 38-site condition). Neither of the potential pretest measures were well balanced either. The ASBs ranged from 0.24 to 0.28 for the academic pretest and from 0.26 to 0.35 for the disciplinary pretest. Therefore, all 11 covariates across treatment conditions, including the academic pretest measure, lacked balance in their distributions, even though the ASBs of some covariates were technically below 0.25. This imbalance of the baseline covariates suggested that the simulation procedures successfully introduced selection bias into the covariates' distributions for all control group size conditions, thus indicating that bias reduction techniques should be used.

4.2.1.2 Post-Adjustment Balance

Tables 7 through 19 show the covariate balance for each of the 13 adjustment methods, where each table represents one specific method. Table 7 presents the covariate balance between treatment conditions after 1:1M.PROG. As shown in Table 7, 1:1M.PROG barely improved the imbalance of covariates other than the pretest measures, as these covariates' adjusted ASBs were approximately equal to their unadjusted ASBs. This result was not surprising because adjustments using PROGs are not supposed to adjust for covariate balance between the treatment and control groups (Wyss, Glynn, & Gagne, 2016). However, this method did produce excellent balance on the pretest measures. The ASBs for the academic pretest were 0.06 under the 3-site without contextual factor condition, 0.04 under the 3-site with contextual factor condition, and 0.00 under

[11] The distribution of "age" across sites was examined as a possible source for the large pre-adjustment imbalance. The range of "age" was not much different from the ranges for other continuous covariates. Therefore, the large pre-adjustment imbalance of "age" did not appear to be related to its range across sites.

the 38-site condition. The ASBs for the disciplinary pretest were 0.06 under the 3-site without contextual factor condition, 0.07 under the 3-site with contextual factor condition, and 0.01 under the 38-site condition. Note that two pretest measures achieved almost perfect balance under the 38-site condition.

Table 8 and Table 9 show the covariate balance between treatment conditions after weighting on the LR-estimated and GBM-estimated PSs, respectively. As shown in these tables, W.PS sufficiently balanced almost all 11 covariates regardless of the PS estimation methods and control group sizes, with the average ASBs across covariates ranging from 0.02 to 0.16 for the LR-estimated PSs and from 0.09 to 0.17 for the GBM-estimated PSs. For weighting on the LR-estimated PSs method, the only unbalanced covariate was "age" under the two 3-site conditions, but its ASB was substantially reduced after adjustment: from 1.03 to 0.35 under the without contextual factor condition and from 1.11 to 0.30 under the with contextual factor condition. Under the 38-site condition, even "age" achieved good balance, with its ASB reduced from 0.88 to 0.04. The largest ASBs among other 10 covariates were 0.17 under either of the two 3-site conditions and 0.06 under the 38-site condition. After adjustment, both pretest measures achieved adequate balance under the two 3-site conditions and good balance under the 38-site condition. The academic pretest had ASBs of 0.12 under the 3-site without contextual factor condition, 0.15 under the 3-site with contextual factor condition, and 0.01 under the 38-site condition. The disciplinary pretest had ASBs of 0.15 under the 3-site without contextual factor condition, 0.17 under the 3-site with contextual factor condition, and 0.06 under the 38-site condition. A similar pattern was found for the GBM-estimated PSs. All covariates were well balanced after adjustment, as the largest ASBs for the covariates were 0.25 under each of the two 3-site conditions and 0.13 under the 38-site condition. In summary, weighting on both the LR-estimated and GBM-estimated PSs

resulted in sufficient balance for almost all covariates. Note that these methods' ability to balance the pretest measure was not as good as that of 1:1M.PROG, particularly under the two 3-site conditions.

Table 10 and Table 11 show the covariate balance between the treatment and control groups after 1:1 optimal matching on the LR- and GBM-estimated PSs, respectively. As shown in Table 10, matching on the LR-estimated PSs resulted in lower post-adjustment than pre-adjustment ASBs for the covariates, leading to an adequate balance for all covariates except for "age" under the two 3-site conditions. Both pretest measures had adequate balance across the control group size conditions, with ASBs ranging from 0.09 to 0.16 for the academic pretest and from 0.11 to 0.16 for the disciplinary pretest. Similar to the LR-estimated PSs, the GBM-estimated PSs also reduced the ASBs of each covariate and yielded sufficient balance for all covariates under the 38-site condition and for eight of the 11 covariates under each of the two 3-site conditions (Table 11). The three covariates that remained unbalanced were "age", "two-parent households", and "had academic risk". The pretest measures also achieved acceptable balance across the control group size conditions after adjustment, with ASBs ranging from 0.13 to 0.21 for the academic pretest and from 0.15 to 0.19 for the disciplinary pretest. Note that, for each covariate, the balance improvement for the GBM-estimated PSs was not as good as it was for the LR-estimated PSs.

Table 12 and Table 13 present the covariate balance between treatment conditions after weighting on the LR- and GBM-estimated ProgPSs, respectively. These two methods had similar performance in balancing covariates. Overall, these two methods barely improved the imbalance of covariates other than the pretest measures, as reflected in the similar average pre- and post-adjustment ASBs across covariates. However, these methods did reduce the ASBs of the pretest measures from close to or above 0.25 to close to 0 across the PS estimation methods and control

group size conditions with only one exception: the ASB of the disciplinary pretest after weighting on the LR-estimated ProgPSs was 0.12 under the 3-site without the contextual factor condition. In other words, the pretest measures after weighting on the LR- or GBM-estimated ProgPSs were well balanced.

Table 14 and 15 present the covariate balance between treatment conditions after matching on the LR- and GBM-estimated ProgPSs, respectively. Like their weighted counterparts, these two methods barely improved the imbalance of covariates other than the pretest measures. However, overall, the pretest measures after 1:1M.ProgPS were not as balanced as those after W.ProgPS. Across the PS estimation methods, the ASBs after 1:1M.ProgPS ranged from 0.07 to 0.19 for the academic pretest and from 0.09 to 0.21 for the disciplinary pretest (Table 14 and Table 15). In contrast, the ASBs after W.ProgPS ranged from 0.00 to 0.04 for the academic pretest and from 0.02 to 0.12 for the disciplinary pretest (Table 12 and Table 13).

Table 16 and Table 17 present the covariate balance between treatment conditions after 1:1 optimal matching on a Mahalanobis distance combining the LR- and GBM-estimated propensity scores and prognostic scores, respectively. Table 18 and Table 19 present the covariate balance between treatment conditions after full matching on a Mahalanobis distance combining the LR- and GBM-estimated propensity scores and prognostic scores, respectively. As shown in Tables 16 – 19, these four methods performed almost identically in terms of balancing covariates. Unlike other adjustments, these four methods led to the increased ASBs for at least nine covariates, including the pretest measures, under each of the two 3-site conditions. As a result, the overall balance under the two 3-site conditions was even worse after adjustment. Under the 38-site condition, these four methods barely improved the imbalance for each covariate, producing similar pre- and post-adjustment ASBs for all covariates.

4.2.2 Summary

In summary, 1:1M.PROG and W.ProgPS produced excellent balance for the pretest measures, particularly under the 38-site condition, even though these two methods did not improve the balance for other covariates. Like 1:1M.PROG and W.ProgPS, 1:1M.ProgPS also yielded adequate balance for the pretest measures but failed to enhance the balance of other covariates. However, 1:1M.ProgPS balanced the pretest measures less effectively than did either 1:1M.PROG or W.ProgPS. Both W.PS and 1:1M.PS yielded adequate balance for almost all of the measured covariates. However, neither balanced the pretest measures as well as 1:1M.PROG or W.ProgPS did. Both had similar results to those of 1:1M.ProgPS. Of all the adjustment methods, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG performed worst in terms of balancing covariates. Under the two 3-site conditions, these two methods worsened the balance for most of the covariates, including the pretest measures. Under the 38-site conditions, they failed to improve the balance for any of the covariates.

The impact of the control group size and the PS estimation methods on the performance of these methods in balancing covariates was also evaluated. I found that, for each method, the covariates were more balanced overall under the 38-site condition than under either of the two 3-site conditions, and I found little difference in covariate balance between the two 3-site conditions. In addition, for methods involving PSs or ProgPSs, their ability to balance covariates was somewhat sensitive to the PS estimation methods. The LR-estimated scores tended to consistently result in slightly better overall covariate balance than that observed for GRM-estimated scores.

4.3 RELATIONSHIP BETWEEN COVARIATE BALANCE AND EFFECT SIZE RECOVERY

The above results suggest that for the matching methods examined, the degree of balance for the pretest measures, rather than the balance for all covariates, had a positive relationship with the effect size estimates. Also, when a method yielded a better balance for the pretest measures, it yielded more precise effect size estimates. The 1:1M.PROG method showed the best performance in recovering the treatment effects and one of the best performances in achieving the pretest balance, although it could not effectively improve the balance of covariates other than the pretest measures. 1:1M.PS and 1:1M.ProgPS were the next two best methods for estimating the treatment effect. Both displayed similar performance in terms of both effect estimation and covariate balance. Unlike 1:1M.PROG, 1:1M.PS and 1:1M.ProgPS resulted in sufficient balance for almost all of the measured covariates. However, 1:1M.PS and 1:1M.ProgPS balanced the pretest measures much less effectively than did the 1:1M.PROG method. The 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG methods exhibited the worst performance of all methods in terms of effect estimation except under the 38-site condition, for which W.PS and W.ProgPS produced even more biased and less efficient estimates. Parallel to their poor performance in effect estimation, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG also displayed the worst performance in balancing covariates, including the pretest measures.

However, for the weighting methods examined (W.PS and W.ProgPS), the degree of balance on the pretest measures provided little information about whether the methods produced efficient and precise effect estimates. Compared to 1:1M.PS, W.PS produced a similarly adequate but slightly superior balance for the pretest measures. However, for effect estimation, W.PS performed no better than the naïve method under the two 3-site conditions. Similarly, W.ProgPS

resulted in excellent balance for the pretest measures, just as 1:1M.PROG did. However, its effect size estimation performance was only somewhat better than the naïve method under the two 3-site conditions. Under the 38-site condition, W.PS and W.ProgPS performed worst of all the methods in terms of effect estimation.

Table 6. A Summary of the Covariate Balance Before Adjustment

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.28	0.24	0.24	0.28	0.24	0.24
Absenteeism rate	0.26	0.35	0.30	0.25	0.34	0.30
Age	1.03	1.11	0.88	1.03	1.11	0.88
Two-parent households	0.33	0.33	0.27	0.33	0.33	0.27
Prior mentoring experience	0.22	0.27	0.22	0.22	0.27	0.22
Had disciplinary risk	0.25	0.23	0.22	0.25	0.23	0.22
Had academic risk	0.55	0.51	0.42	0.55	0.51	0.42
Female	0.30	0.29	0.25	0.31	0.29	0.25
GPA	0.30	0.31	0.29	0.30	0.31	0.29
Pro-social behaviors	0.27	0.31	0.22	0.27	0.26	0.22
Future orientation	0.35	0.36	0.28	0.36	0.36	0.28
Average	0.38	0.39	0.33	0.38	0.39	0.33

Note. Covariate balance measure is absolute standardized bias.

Table 7. A Summary of the Covariate Balance after 1:1M.PROG

Covariates	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.06	0.04	0.00	0.34	0.28	0.26
Absenteeism rate	0.28	0.35	0.34	0.06	0.07	0.01
Age	0.94	0.99	0.83	1.00	1.03	0.91
Two-parent households	0.33	0.33	0.26	0.33	0.34	0.27
Prior mentoring experience	0.25	0.24	0.24	0.24	0.28	0.22
Had disciplinary risk	0.21	0.24	0.19	0.25	0.22	0.23
Had academic risk	0.57	0.50	0.41	0.56	0.49	0.40
Female	0.28	0.35	0.25	0.30	0.33	0.26
GPA	0.28	0.35	0.29	0.35	0.31	0.26
Pro-social behaviors	0.22	0.21	0.20	0.29	0.28	0.29
Future orientation	0.32	0.38	0.21	0.35	0.35	0.29
Average	0.34	0.36	0.29	0.37	0.36	0.31

Note. 1:1M.PROG = 1:1 optimal matching on the estimated prognostic scores. Covariate balance measure is absolute standardized bias.

Table 8. A Summary of the Covariate Balance after W.LR.PS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.12	0.15	0.01	0.12	0.15	0.01
Absenteeism rate	0.15	0.16	0.06	0.15	0.17	0.06
Age	0.35	0.30	0.04	0.36	0.30	0.04
Two-parent households	0.16	0.14	0.01	0.16	0.15	0.01
Prior mentoring experience	0.15	0.13	0.01	0.15	0.13	0.01
Had disciplinary risk	0.13	0.16	0.02	0.14	0.16	0.02
Had academic risk	0.10	0.09	0.01	0.11	0.10	0.01
Female	0.13	0.15	0.01	0.13	0.15	0.01
GPA	0.16	0.16	0.01	0.16	0.16	0.01
Pro-social behaviors	0.13	0.13	0.01	0.13	0.13	0.01
Future orientation	0.12	0.13	0.02	0.12	0.13	0.02
Average	0.15	0.15	0.02	0.16	0.16	0.02

Note. W.LR.PS = Weighting on the logistic regression estimated propensity scores. Covariate balance measure is absolute standardized bias.

Table 9. A Summary of the Covariate Balance after W.GBM.PS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.15	0.12	0.06	0.15	0.12	0.06
Absenteeism rate	0.14	0.12	0.08	0.14	0.12	0.08
Age	0.25	0.24	0.13	0.25	0.24	0.13
Two-parent households	0.19	0.21	0.10	0.19	0.22	0.10
Prior mentoring experience	0.17	0.14	0.08	0.17	0.14	0.08
Had disciplinary risk	0.13	0.14	0.09	0.13	0.14	0.09
Had academic risk	0.23	0.25	0.12	0.23	0.25	0.12
Female	0.15	0.16	0.08	0.15	0.16	0.08
GPA	0.16	0.14	0.08	0.16	0.14	0.08
Pro-social behaviors	0.13	0.16	0.07	0.13	0.16	0.07
Future orientation	0.20	0.20	0.10	0.20	0.20	0.10
Average	0.17	0.17	0.09	0.17	0.17	0.09

Note. W.GBM.PS = Weighting on the generalized boosted models estimated propensity scores. Covariate balance measure is absolute standardized bias.

Table 10. A Summary of the Covariate Balance after 1:1M.LR.PS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.16	0.14	0.09	0.16	0.14	0.09
Absenteeism rate	0.15	0.16	0.11	0.15	0.16	0.11
Age	0.44	0.39	0.12	0.44	0.39	0.12
Two-parent households	0.15	0.17	0.10	0.15	0.17	0.10
Prior mentoring experience	0.12	0.13	0.11	0.12	0.13	0.11
Had disciplinary risk	0.12	0.14	0.09	0.12	0.14	0.09
Had academic risk	0.23	0.22	0.10	0.23	0.22	0.10
Female	0.11	0.14	0.10	0.11	0.14	0.10
GPA	0.16	0.16	0.11	0.16	0.16	0.11
Pro-social behaviors	0.18	0.14	0.11	0.18	0.14	0.11
Future orientation	0.16	0.15	0.12	0.16	0.15	0.12
Average	0.18	0.18	0.11	0.18	0.18	0.11

Note. 1:1M.LR.PS = 1:1 optimal matching on the logistic regression estimated propensity scores. Covariate balance measure is absolute standardized bias.

Table 11. A Summary of the Covariate Balance after 1:1M.GBM.PS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.21	0.17	0.13	0.21	0.17	0.13
Absenteeism rate	0.18	0.19	0.15	0.18	0.19	0.15
Age	0.44	0.40	0.19	0.44	0.40	0.19
Two-parent households	0.26	0.27	0.10	0.26	0.27	0.10
Prior mentoring experience	0.17	0.19	0.11	0.17	0.19	0.11
Had disciplinary risk	0.15	0.15	0.13	0.15	0.15	0.13
Had academic risk	0.37	0.27	0.13	0.37	0.28	0.13
Female	0.17	0.18	0.12	0.17	0.19	0.12
GPA	0.21	0.21	0.11	0.21	0.20	0.11
Pro-social behaviors	0.20	0.20	0.18	0.20	0.20	0.18
Future orientation	0.22	0.20	0.15	0.22	0.21	0.16
Average	0.23	0.22	0.14	0.23	0.22	0.14

Note. 1:1M.GBM.PS = 1:1 optimal matching on the generalized boosted models estimated propensity scores. Covariate balance measure is absolute standardized bias.

Table 12. A Summary of the Covariate Balance after W.LR.ProgPS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.04	0.04	0.00	0.26	0.24	0.24
Absenteeism rate	0.29	0.33	0.31	0.05	0.12	0.04
Age	0.84	0.87	0.79	0.96	1.01	0.84
Two-parent households	0.34	0.31	0.27	0.33	0.34	0.27
Prior mentoring experience	0.23	0.27	0.23	0.21	0.26	0.20
Had disciplinary risk	0.22	0.22	0.19	0.22	0.23	0.24
Had academic risk	0.46	0.48	0.38	0.47	0.45	0.34
Female	0.29	0.30	0.23	0.34	0.33	0.25
GPA	0.25	0.28	0.25	0.30	0.27	0.23
Pro-social behaviors	0.19	0.22	0.16	0.26	0.27	0.22
Future orientation	0.32	0.32	0.23	0.39	0.32	0.28
Average	0.32	0.33	0.28	0.34	0.35	0.29

Note. W.LR.ProgPS = Weighting on the logistic regression estimated prognostic propensity scores. Covariate balance measure is absolute standardized bias.

Table 13. A Summary of the Covariate Balance after W.GBM.ProgPS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.03	0.02	0.01	0.28	0.28	0.25
Absenteeism rate	0.27	0.35	0.29	0.04	0.03	0.02
Age	0.99	0.99	0.80	1.04	1.08	0.87
Two-parent households	0.32	0.35	0.26	0.32	0.34	0.27
Prior mentoring experience	0.25	0.28	0.22	0.22	0.27	0.22
Had disciplinary risk	0.23	0.22	0.19	0.23	0.21	0.21
Had academic risk	0.54	0.50	0.40	0.53	0.50	0.38
Female	0.31	0.32	0.24	0.29	0.31	0.27
GPA	0.30	0.31	0.26	0.34	0.27	0.27
Pro-social behaviors	0.20	0.22	0.16	0.27	0.26	0.23
Future orientation	0.31	0.38	0.24	0.37	0.36	0.25
Average	0.34	0.36	0.28	0.36	0.36	0.29

Note. W.GBM.ProgPS = Weighting on the generalized boosted models estimated prognostic propensity scores. Covariate balance measure is absolute standardized bias.

Table 14. A Summary of the Covariate Balance after 1:1M.LR.ProgPS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.14	0.10	0.07	0.31	0.27	0.27
Absenteeism rate	0.29	0.34	0.35	0.10	0.12	0.09
Age	0.91	0.87	0.83	1.01	0.99	0.83
Two-parent households	0.32	0.34	0.25	0.32	0.36	0.28
Prior mentoring experience	0.23	0.24	0.24	0.20	0.27	0.22
Had disciplinary risk	0.23	0.24	0.24	0.24	0.21	0.29
Had academic risk	0.51	0.46	0.41	0.51	0.44	0.35
Female	0.29	0.30	0.27	0.33	0.31	0.32
GPA	0.29	0.33	0.26	0.27	0.28	0.28
Pro-social behaviors	0.26	0.25	0.21	0.30	0.28	0.24
Future orientation	0.34	0.30	0.25	0.32	0.26	0.29
Average	0.35	0.35	0.31	0.36	0.34	0.31

Note. 1:1M.LR.ProgPS = 1:1 optimal matching on the logistic regression estimated prognostic propensity scores. Covariate balance measure is absolute standardized bias.

Table 15. A Summary of the Covariate Balance after 1:1M.GBM.ProgPS

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.19	0.16	0.15	0.31	0.31	0.30
Absenteeism rate	0.28	0.35	0.33	0.20	0.21	0.21
Age	1.02	0.88	0.81	1.04	1.03	0.88
Two-parent households	0.36	0.34	0.31	0.34	0.34	0.26
Prior mentoring experience	0.24	0.27	0.27	0.21	0.25	0.21
Had disciplinary risk	0.25	0.23	0.24	0.23	0.21	0.21
Had academic risk	0.55	0.42	0.37	0.53	0.47	0.35
Female	0.30	0.33	0.27	0.28	0.33	0.28
GPA	0.34	0.34	0.29	0.34	0.32	0.30
Pro-social behaviors	0.26	0.26	0.20	0.30	0.29	0.29
Future orientation	0.36	0.35	0.24	0.37	0.40	0.28
Average	0.38	0.36	0.32	0.38	0.38	0.32

Note. 1:1M.GBM.ProgPS = 1:1 optimal matching on the generalized boosted models estimated prognostic propensity scores. Covariate balance measure is absolute standardized bias.

Table 16. A Summary of the Covariate Balance after 1:1MAHL.LRPS.PROG

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.34	0.27	0.22	0.35	0.28	0.22
Absenteeism rate	0.28	0.45	0.28	0.28	0.43	0.28
Age	1.08	1.09	0.95	1.09	1.10	0.94
Two-parent households	0.33	0.29	0.27	0.33	0.29	0.27
Prior mentoring experience	0.30	0.25	0.25	0.30	0.25	0.25
Had disciplinary risk	0.26	0.27	0.27	0.28	0.28	0.26
Had academic risk	0.52	0.51	0.39	0.51	0.50	0.39
Female	0.31	0.33	0.29	0.32	0.33	0.29
GPA	0.37	0.39	0.30	0.37	0.38	0.30
Pro-social behaviors	0.34	0.36	0.22	0.35	0.36	0.22
Future orientation	0.44	0.35	0.26	0.45	0.36	0.26
Average	0.41	0.41	0.33	0.42	0.42	0.34

Note. 1:1MAHL.LRPS.PROG = 1:1 optimal matching on a Mahalanobis distance combining the logistic regression estimated propensity and prognostic scores. Covariate balance measure is absolute standardized bias.

Table 17. A Summary of the Covariate Balance after 1:1MAHL.GBMPS.PROG

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.35	0.31	0.23	0.36	0.30	0.24
Absenteeism rate	0.31	0.47	0.28	0.30	0.46	0.28
Age	1.11	1.10	0.96	1.13	1.18	0.95
Two-parent households	0.35	0.32	0.28	0.34	0.32	0.28
Prior mentoring experience	0.29	0.29	0.28	0.29	0.28	0.29
Had disciplinary risk	0.27	0.26	0.26	0.28	0.26	0.25
Had academic risk	0.53	0.44	0.40	0.53	0.46	0.40
Female	0.31	0.32	0.30	0.33	0.31	0.30
GPA	0.36	0.38	0.29	0.36	0.38	0.29
Pro-social behaviors	0.36	0.36	0.22	0.37	0.36	0.22
Future orientation	0.44	0.35	0.25	0.44	0.35	0.26
Average	0.43	0.42	0.34	0.43	0.42	0.34

Note. 1:1MAHL.GBMPS.PROG = 1:1 optimal matching on a Mahalanobis distance combining the generalized boosted models estimated propensity and prognostic scores. Covariate balance measure is absolute standardized bias.

Table 18. A Summary of the Covariate Balance after FULL.MAHL.LRPS.PROG

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.35	0.28	0.23	0.36	0.29	0.23
Absenteeism rate	0.27	0.46	0.28	0.27	0.45	0.28
Age	1.12	1.12	0.96	1.12	1.14	0.95
Two-parent households	0.33	0.29	0.27	0.33	0.29	0.27
Prior mentoring experience	0.31	0.25	0.25	0.31	0.25	0.25
Had disciplinary risk	0.28	0.28	0.26	0.29	0.28	0.26
Had academic risk	0.50	0.51	0.40	0.49	0.50	0.40
Female	0.33	0.33	0.30	0.34	0.34	0.30
GPA	0.37	0.39	0.30	0.38	0.38	0.30
Pro-social behaviors	0.34	0.36	0.22	0.34	0.36	0.22
Future orientation	0.45	0.35	0.26	0.45	0.35	0.26
Average	0.42	0.42	0.34	0.43	0.42	0.34

Note. FULL.MAHL.LRPS.PROG = Full matching on a Mahalanobis distance combining the logistic regression estimated propensity and prognostic scores. Covariate balance measure is absolute standardized bias.

Table 19. A Summary of the Covariate Balance after FULL.MAHL.GBMPS.PROG

	Size of Control Groups (Number of Sites)					
	Academic Outcome			Disciplinary Outcome		
Covariates	3 (without contextual factor)	3 (with contextual factor)	38	3 (without contextual factor)	3 (with contextual factor)	38
Student efficacy and bonding	0.36	0.30	0.24	0.37	0.29	0.24
Absenteeism rate	0.31	0.47	0.28	0.31	0.47	0.28
Age	1.13	1.11	0.97	1.14	1.23	0.97
Two-parent households	0.35	0.31	0.28	0.34	0.31	0.28
Prior mentoring experience	0.30	0.29	0.28	0.31	0.29	0.28
Had disciplinary risk	0.29	0.26	0.25	0.30	0.26	0.25
Had academic risk	0.52	0.44	0.40	0.52	0.48	0.40
Female	0.33	0.31	0.30	0.34	0.32	0.30
GPA	0.35	0.37	0.29	0.36	0.38	0.29
Pro-social behaviors	0.36	0.36	0.22	0.36	0.36	0.22
Future orientation	0.44	0.35	0.26	0.45	0.36	0.26
Average	0.43	0.42	0.34	0.44	0.43	0.34

Note. FULL.MAHL.GBMPS.PROG = Full matching on a Mahalanobis distance combining the generalized boosted models estimated propensity and prognostic scores. Covariate balance measure is absolute standardized bias.

5.0 DISCUSSION

Despite the increased emphasis on randomized controlled trials (RCTs) in the field of education over the past decade (Shadish & Cook, 2009), opportunities for true RCTs may be limited due to practical and ethical concerns. As a result, many education researchers have had to rely on observational data to answer causal questions. Because observational studies are often vulnerable to selection bias, sound observational methods are desirable to effectively reduce or remove selection bias. Two summary score methods, the propensity score (PS) and the prognostic score (PROG), have become increasingly advocated for controlling selection bias in nonexperimental studies. In addition, a promising new adjustment method—the joint use of propensity and prognostic scores (the two-score method)—has been proposed to improve the performance of the PS adjustment. Because this is a new method, its utility in controlling for bias has not been thoroughly investigated.

Therefore, the primary goal of this study was to investigate whether the joint use of PSs and PROGs is more effective than the use of PSs or PROGs alone for reducing bias in a real educational evaluation context. To this end, a simulation study was conducted to compare the two-score method with the single score methods. Recovery of a simulated treatment effect was compared under experimental conditions based on different control group sizes, outcome measures, and propensity score estimation methods.

In the following sections of this chapter, I first summarize the study's major findings and discuss the implications of these findings for methodologists and applied researchers in the field.

Then I discuss the study's limitations and provide suggestions for potential areas of future research. Lastly, I complete this chapter with a conclusion.

5.1 MAJOR FINDINGS AND IMPLICATIONS

5.1.1 Two-Score Method – Combining Propensity and Prognostic Scores

One of this study's major findings was that the simulation results do not support the use of the two-score methods to estimate the average treatment effect on the treated (ATT). Four two-score methods were evaluated in this study: 1:1 optimal matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (1:1MAHAL.PS.PROG), full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores (FULL.MAHAL.PS.PROG), weighting on the estimated prognostic propensity scores (W.ProgPS), and 1:1 optimal matching on the estimated prognostic propensity scores (1:1M.ProgPS). In terms of the effect recovery, the first three methods performed no better than a simple t-test (i.e., no adjustment with any covariates) across all simulation conditions. The 1:1M.ProgPS method outperformed a simple t-test, but its performance was highly comparable to that of 1:1 matching on the estimated propensity scores (1:1M.PS), and both of these methods performed worse than 1:1 matching on the estimated prognostic scores (1:1M.PROG).

The findings regarding the performances of 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG were not consistent with the findings of the prior simulation research (Hansen, 2006; Leacy & Stuart, 2014). The results of these two simulation studies showed that FULL.MAHAL.PS.PROG was a promising two-score method in confounding control. In Hansen

(2006), FULL.MAHAL.PS.PROG, with or without the PS or PROG caliper, yielded less biased and more efficient effect estimates than full matching on the PSs alone. Leacy and Stuart (2014) also found that compared to other two-score and single-score adjustments, FULL.MAHAL.PS.PROG displayed the strongest robustness in effect estimation across various scenarios of model misspecifications. In my simulations, however, FULL.MAHAL.PS.PROG and 1:1MAHAL.PS.PROG exhibited the worst performance of all methods under the two 3-site conditions and were only better than two weighting adjustments under the 38-site condition.

The reason that 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG performed so poorly may relate to the non-normal distribution of the estimated summary scores. As discussed in Chapter 2, Mahalanobis distance matching works well when there are relatively few (less than five) covariates and when those covariates are approximately normally distributed (Gu & Rosenbaum, 1993; Rubin & Thomas, 2000). The 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG methods have only two matching variables each (i.e., PROGs and linear PSs), so non-normal score distributions may partially account for the poor performance of these two methods. To verify this, I examined the distributions of the linear PSs and PROGs in each crossed condition for several replications. I found that the score distributions were not normal in most cases and that there were greater numbers of non-normal distributions for the control group than for the treatment group, as the former had more outliers in the distributions.

The results also revealed that 1:1M.ProgPS had no advantage when compared to 1:1M.PS and that it was in a clear disadvantage when compared to 1:1M.PROG, even though 1:1M.ProgPS was the best performing two-score method evaluated. This finding was also inconsistent with those of the previous simulation studies (Leacy & Stuart, 2014; Tu & Koh, 2017). In Leacy and Stuart (2014), full matching on the ProgPSs within the PS caliper exhibited better performances than the

PS or PROG adjustments in most cases, and particularly when both the PS and PROG models were incorrectly specified. Tu and Koh (2017) investigated the comparative performance of ProgPSs (either predicted by PROGs only, or predicted by PROGs along with other covariates), PROGs, and PSs when these four summary scores were paired with the 1:1 nearest neighbor matching within varying caliper widths. These researchers found that the performances of these four summary scores were similar in terms of bias and mean squared error. According to these findings, the caliper is probably the main factor in the cross-study performance differences between matching on the ProgPSs and matching on PSs or PROGs. Both Leacy and Stuart (2014) and Tu and Koh (2017) performed matching on ProgPSs within the PS caliper, but I implemented it without a caliper because I aimed to maintain the same effect estimand across all methods. As a result, some of the matched pairs in this study may have poor quality because they lacked the restrictions of the caliper. Furthermore, these differences may also be partly due to the studies' simulation settings. For example, Leacy and Stuart (2014) and Tu and Koh (2017) placed their simulations in a medical context, using artificial data and large sample sizes (N = 1000, 2000, and 5000 in Leacy & Stuart, 2014; N = 5000 in Tu & Koh, 2017), whereas my simulations were in an educational context, using real data (except for the outcomes) and smaller sample sizes (N < 400 under the two 3-site conditions).

In addition, this study's results also do not support the use of W.ProgPS as a two-score method for estimating ATT. The W.ProgPS method and the weighting on the estimated propensity scores (W.PS) method exhibited the worst performances of all methods under the 38-site condition, as they produced both extremely large biases and root mean squared deviations (RMSDs). They also performed poorly under the two 3-site conditions. I discuss the possible reasons for the poor performance of these two weighting adjustments in a later section.

In summary, the findings of this study do not provide evidence to support using Hansen's (2006, 2008) two-score methods to control for selection bias in observational studies with conditions similar to this study's. Either the conditions or the implementation of the two-score methods in my simulations may have caused this inconsistency. Thus, more research is needed to investigate the conditions under which the two-score methods may be preferable to single score adjustments. More importantly, there is still not much research on the strategies of combining PSs and PROGs. In addition to Hansen's methods, Leacy and Stuart (2014) proposed combining both scores through joint subclassification, but this combination method exhibited much worse performance than either of Hansen's two methods. Kelcey and Swoboda (2015) suggested the use of PS matching within PROG strata, but this method's performance has not yet been evaluated in a simulation study. Thus, methodologists should investigate alternative strategies for integrating the propensity and prognostic scores.

5.1.2 Adjustment Based on Prognostic Scores

In this study, 1:1M.PROG was found to exhibit lower RMSD in recovering the simulated treatment effect than any of the two-score and PS adjustments across all simulation conditions. Leacy and Stuart (2014) reported similar findings: regardless of whether the PS model is correctly specified, full matching on the estimated PROGs may be preferred to the two-score and other single-score methods as long as the PROG model is correctly specified. In Tu and Koh (2017), matching on PROGs exhibited comparable performance to matching on PSs or ProgPSs when both the PS and PROG models were correctly specified. Unlike in Leacy and Stuart (2014) and in Tu and Koh (2017), the PROG model was incorrectly specified in this study because of the inconsistencies between the true outcome generating model (Model 7) and the PROG estimation

model (Model 9). Moreover, the PS and ProgPS models were also likely incorrectly specified in my study (as their true models were unknown). However, despite the presence of these misspecified summary score models, 1:1M.PROG still outperformed all other methods. These results suggest that matching on PROGs may still be preferable to two-score adjustments or adjustments using PSs alone when the PROG model contains the prognostically important covariates and when the PS model contains many prognostically weak covariates.

Although matching on PROGs appears to be a promising alternative to the PS adjustment for effect estimation when the PROG model can be correctly specified, researchers need to be aware of the challenges found in modeling PROGs. First, strategies for validating the PROG models are not well developed. The estimated PROGs are susceptible to various types of model misspecification, including overfitting, omitted covariates or high-order terms in the functional form, and the dimensionality of PROG (i.e., a scalar vs. a vector-valued function for continuous outcomes). In practice, researchers commonly use goodness-of-fit statistics and prediction diagnostics to validate the PROG models. However, the validity of the fitted PROG models should be evaluated through checks on the prognostic balance (Hansen, 2008).

As discussed in Chapter 2, unlike the propensity balance, the prognostic balance can only be evaluated within the control group, as the potential outcomes of the treatment group members in the control condition cannot be observed. Some researchers have proposed a “dry-run” analysis to evaluate the prognostic balance for the entire sample (Hansen, 2006; Wyss et al., 2017). This method uses the estimated PSs to divide the control population into “pseudo-treatment” and “pseudo-control” groups, with the differences between these two groups reflecting those between the treatment and control groups in the original sample. The analysis then involves fitting the PROG model to the pseudo-control group. Thus, the prognostic balance can be evaluated within

the entire pseudo-population. Detailed steps for performing a dry-run analysis can be found in Hansen (2006) and in Wyss et al. (2017). Although the dry-run analysis performs well in certain settings, it is limited by its dependence on the accurate estimation of PSs because it uses the estimated PSs to create the pseudo-population (Wyss et al., 2017). Further, performing a dry-run analysis can be very challenging, given its technical complexity. Thus, researchers should focus on developing an effective diagnosis tool for PROG model misspecification.

Second, modeling PROGs requires the use of observed outcomes, even though the outcomes for only one group (usually the control group) are used. A key advantage of the PS adjustment is that it does not use the outcomes in the PS estimation process, so it has a clear separation of "design" and "analysis (Rubin, 2007). This advantage even allows the estimated PSs to serve as a handy tool to create matched groups in the design stage of a prospective study before the outcomes are collected (Stuart, 2010). For this reason, some researchers may question whether the use of the PROGs reduces a study's objectivity and the validity of its inferences.

Note that in Leacy and Stuart (2014), another PROG adjustment—subclassification on the estimated PROGs—performed worse than all the two-score methods and full matching on the PROGs across all settings. Only one PROG adjustment—matching on PROGs—was evaluated in this study and in Tu and Koh (2017). Hence, future work should compare the performances of stratification and matching on PROGs to ascertain whether matching on PROGs truly has the advantage.

5.1.3 Adjustment Based on Propensity Scores

Two PS adjustments (1:1M.PS and W.PS) were examined in this study. The simulation results revealed that, in terms of effect size recovery, matching on PSs had a clear advantage over

its weighting counterpart and was not at a disadvantage relative to any of the two-score methods examined in the study. More specifically, 1:1M.PS and 1:1M.ProgPS had highly comparable results, and both methods were inferior only to a prognostic score adjustment, 1:1M.PROG. As discussed in the previous sections, more research on PROGs and ProgPSs is needed to support their uses in practical applications. Thus, in practice, matching on PSs may be a viable alternative to matching on PROGs or ProgPSs.

Unlike 1:1M.PS, W.PS did not display a satisfactory performance in recovering the simulated treatment effect. Regardless of the PS estimation methods, W.PS and W.ProgPS exhibited the worst performance of all the adjustment methods under the 38-site condition. Under the two 3-site conditions, W.PS performed better than only 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG. Note that, even though both W.PS and W.ProgPS performed poorly across all conditions, W.ProgPS still performed slightly better than did W.PS, particularly when PSs or ProgPSs were estimated using GBM. Further detailed discussions about W.PS are provided in the next section (weighting adjustments).

5.1.4 Weighting Adjustments

One unexpected finding from this study was that, compared to other methods, the weighting adjustments evaluated (W.PS and W.ProgPS) showed a strong tendency to yield overestimated ATT estimates with large variability, particularly under the 38-site condition. One possible explanation for the large bias and variability of the weighting estimators may be the extremely low weights observed in some control group students. As a check, the distribution of the weights for the control group students across all 39 replications was examined (see Table 20). The PS- and ProgPS-derived weights were very low for most of the control group students across

all outcome measures and all PS estimation methods. This was the case in particular for the 38-site condition, where the maximum weights for 99% of control group students ranged from 0.08 to 0.19 depending on the outcomes and PS/ProgPS estimation methods. Recall that, for ATT weighting, all treatment group members receive a weight of 1 and control group members receive a weight of $\frac{PS}{1-PS}$, and the formula used for calculating effect size is $d = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{(n_T-1)SD_T^2+(n_C-1)SD_C^2}{n_T+n_C-2}}}$ and $SD_{weight}^2 = \frac{1}{n-1}\sum w_i\,(x_i - \bar{X}_{weight})^2$. Given these formulas, low weights, when combined with large n_C values under the 38-site condition, tended to produce very large effect size estimates, which in turn resulted in extremely high values for bias and RMSD under that condition.

In addition to extremely low weights, extremely large weights may also have contributed to large RMSDs under the 3-site conditions. Researchers have suggested that observations with extremely large weights may improperly influence results, thus yielding estimates with high variance (Rubin, 2001; Schafer & Kang, 2008). As shown in Table 20, under the two 3-site conditions, the LR-estimated PSs for both academic and disciplinary outcomes, as well as the LR-estimated ProgPSs for the disciplinary outcome, produced some extremely large weights despite a very small percentage of such weights (less than 1%). Here, weights greater than 10 are considered to be extremely large, following Harder et al. (2010) and Lee et al. (2010). Overall, these extreme weights may be due to PS or ProgPS model misspecification. According to Rubin (2001), weighting is more sensitive than matching and stratification to the misspecification of the PS (including the ProgPS) models because weighting directly uses PSs (including ProgPSs) to estimate the treatment effect. In this study, misspecifications of the PS and ProgPS models could explain the extreme weights, as the correct PS and ProgPS models were unknown.

Obviously, these weighting results echo other researchers' claims about the PS weighting estimators. That is, the PS weighting estimators can be biased and inefficient due to the extreme weights (e.g., Rubin, 2001; Schafer & Kang, 2008). More importantly, this study's results suggest that weighting estimators based on ProgPSs or PSs have similar properties. To the best of my knowledge, this study is the first to examine the performances of W.ProgPS in estimating ATT. Therefore, methodologically, this study provides an important addition to the understanding of the utility of the two-score methods.

These findings also indicate that applied researchers need to be cautious when applying weighting applications based on PSs or ProgPSs for ATT estimation. It is true that, compared to matching, weighting approaches have the advantage of keeping all the samples. However, the weighting estimators are more likely to be biased and inefficient due to the model misspecification. This could discredit the inferences that are obtained from the results and result in some unwanted consequences. For example, the highly inflated weighting estimators as those in this study could lead the researchers to believe that there is a strong, educationally meaningful treatment effect that otherwise does not exist. This, in turn, may cause those who create policy to make wrong decisions.

Table 20. Distribution of PS-/ProgPS-Based Weights for the Control Group by the Estimation Method

Outcome Measures	Summary Score	Estimation Method	Size of Control Groups (Number of Sites)	*N*	Mean	1st quartile	Median	3rd quartile	99%	Maximum
Academic / disciplinary	PS	LR	3 (with contextual factor)	5609	0.39	0.03	0.10	0.31	3.23	274.06
			3 (without contextual factor)	5162	0.40	0.04	0.12	0.34	3.47	154.02
			38	68476	0.03	0.00	0.01	0.03	0.19	6.14
		GBM	3 (with contextual factor)	5609	0.06	0.00	0.01	0.04	0.76	3.10
			3 (without contextual factor)	5162	0.10	0.00	0.02	0.10	1.05	4.66
			38	68476	0.01	0.00	0.00	0.01	0.13	1.59
Academic	ProgPS	LR	3 (with contextual factor)	5609	0.32	0.14	0.25	0.41	1.31	8.67
			3 (without contextual factor)	5162	0.35	0.15	0.26	0.43	1.54	7.59
			38	68476	0.03	0.01	0.02	0.03	0.09	0.61
		GBM	3 (with contextual factor)	5609	0.23	0.07	0.14	0.29	1.23	4.30
			3 (without contextual factor)	5162	0.26	0.09	0.18	0.32	1.30	5.89
			38	68476	0.02	0.01	0.01	0.03	0.15	1.06
Disciplinary	ProgPS	LR	3 (with contextual factor)	5609	0.33	0.14	0.26	0.43	1.35	17.69
			3 (without contextual factor)	5162	0.37	0.18	0.27	0.41	1.39	98.01
			38	68476	0.03	0.01	0.02	0.03	0.08	2.13
		GBM	3 (with contextual factor)	5609	0.25	0.07	0.18	0.33	1.29	5.59
			3 (without contextual factor)	5162	0.25	0.07	0.17	0.32	1.42	7.73
			38	68476	0.02	0.00	0.01	0.02	0.18	1.63

Note. PS = Propensity score; ProgPS = Prognostic propensity score; LR = Logistic regression; GBM = Generalized boosted models.

5.1.5 Pretest Balance

Another finding from this study was that matching methods with better balance on the pretest measures (as measured by the absolute standardized bias across treatment conditions) demonstrated more precise effect estimates. For example, 1:1M.PROG produced the lowest RMSD even though it failed to improve the balance of the covariates other than the pretest measures. Conversely, 1:1M.PS and 1:1M.ProgPS produced relatively larger RMSDs even though they had two of the lowest average balance across covariates. Similarly, 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG exhibited the worst performance of all matching methods in terms of both recovering effect sizes and balancing covariates (including the pretest measures). These results support the findings of Cham (2013) and Lee et al. (2010), who demonstrated that PSs that produced the best average balance across covariates did not necessarily produce the least biased effect estimates. This study's results suggest that the prior findings also apply to the adjustments based on PROGs or on both scores. Furthermore, this study's results found that a higher degree of pretest balance (pretest was the only prognostically important covariate in this study), not a greater average balance across covariates, corresponded with better confounding control for all the two-score and one-score matching methods.

These findings have important implications for constructing PS (including ProgPS) or PROG models and/or for selecting the appropriate method to utilize these summary scores in practice. First, pretest measures of outcomes (if available) should always be included in summary score models. Although it is actually important to include any prognostically important covariates in the score estimation model, in practice, researchers may not know which covariates are

prognostically important except for the pretest measures. This is especially the case for PS adjustments as the outcomes are not supposed to be used for PS building. One limitation of PS is its handling of prognostically weak covariates in the PS estimation. Covariates with the same relationship to treatment assignment are treated the same regardless of their relationship with the outcomes (Rubin, 1997). However, inclusion of prognostically weak covariates reduces the efficiency and sometime increase the bias of the effect estimates (e.g., Brookhart et al., 2006). Therefore, pretest measures should be favored when constructing the summary score models.

Second, a score estimation-application combination that can maximize the degree of balance on the pretest measures should be selected. Scholars have considered covariate balance as a key criterion in determining the validity of the estimated PSs (Austin, 2011). Thus, in practical applications, analysts routinely select a PS model specification and an adjustment algorithm based on their ability to balance the measured covariates (Austin, 2011). For example, researchers generally tend to select methods that result in the lowest average balance across all measured covariates, or the fewest number of covariates with large differences. This study's results indicated that balancing the pretest measures was more critical than balancing other covariates to obtain more accurate effect estimates. Based on these results, applied researchers should select the score estimation-application combination that yields the highest degree of balance on the covariates that are most predictive of the outcomes.

This recommendation is not only limited to the PS adjustments. As noted earlier, measures to evaluate the prognostic balance are not readily available. In this case, the balance of the pretest measures (or other prognostically important measures) across treatment groups seems to be a convenient diagnostic for assessing the adequacy of a PROG specification. In other words, for any adjustment using summary scores, if enough balance cannot be achieved on the prognostically

important covariates in a particular data set, the credibility of the treatment effect's inferences may be questionable. In sum, a tentative conclusion based on this study's results is that checking the pretest balance can help detect which adjustment methods will produce a very biased estimate of an ATT effect.

Note that, even though this study appears to suggest that the pretest balance is quite sensitive to detect adjustment methods that produce less biased versus more biased effect estimates, researchers should not assume that the balance for pretest measures (or even all the covariates that are presumed to be prognostically important) is sufficient to ensure accurate effect estimates. Furthermore, the above recommendations may only work for the matching adjustments. For weighting procedures, the degree of balance on the prognostic covariates after adjustment provides little information about which method would produce more accurate effect estimates. For example, W.ProgPS and 1:1M.PROG resulted in equally good balance for the pretest of the outcome measure, but the former method's effect estimates were much more biased and inefficient than the latter's. In addition, similar to 1:1M.PS, W.PS resulted in an acceptable balance on all 11 covariates. However, its performance was even worse than that of W.ProgPS. Further investigation into the relationship between pretest balance and the accuracy of the weighting estimators is needed.

5.1.6 Factors Influencing Method Performance

In addition to the differences in method performance, this study also investigated which factors influenced the performances of these methods. Three factors were examined: PS estimation method, control group size, and outcome measures. Researchers have found that these three factors

affect the performance of the PS methods. However, no scholars have yet evaluated their influence on the performance of the two-score methods. Thus, the topic of this investigation is of interest to both methodologists and applied researchers.

The first factor evaluated was the PS estimation method. Specifically, I was interested in whether the performance of the two-score and one-score adjustment methods using PSs or ProgPSs would be sensitive to two PS estimation methods: generalized boosted modeling (GBM) and logistic regression (LR). The simulation results indicated that GBM did not have any advantage relative to LR. For 1:1MAHAL.PS.PROG and FULL.MAHAL.PS.PROG, GBM and LR produced highly comparable results in terms of covariate balance and effect estimation. On the other hand, for other methods involving PSs or ProgPSs, compared to LR, GBM tended to yield slightly greater RMSD and slightly worse covariate balance across other conditions, particularly for weighting adjustments. These findings were somewhat inconsistent with those of the prior research. Researchers have found that GBM tends to perform well when paired with weighting (e.g., Harder et al., 2010; Lee et al., 2009) but it exhibits poor performance when paired with matching (e.g., Diamond & Sekhon, 2013; Pirracchio et al., 2015; Stone & Tang, 2013).

The reason for GBM's relatively inferior performance with matching may be that GBM was originally developed specifically for weighting applications. However, in this study, GBM tended to perform slightly worse than LR for both weighting and matching adjustments (other than for two adjustments involving Mahalanobis distance matching). One possible explanation for GBM's inferior performance is the lack of overlap between the treatment and control groups in the GBM-estimated scores. Compared to LR, GBM appeared to result in a distribution of scores with less overlap between the treatment and control groups. However, it is not clear why GBM tended to yield distributions with less overlap, so this topic still needs further study.

In addition, these results suggest that, in certain settings, LR is preferable to GBM despite the theoretical advantages of the latter method. First, the GBM-estimated PSs or ProgPSs might not exhibit their best performance in a study with a limited number of covariates. A major disadvantage of LR is its requirement for using iterative specification of the PS model to investigate not only covariate main effects but also higher order interactions. This process can be time-consuming, and there is no guarantee of success, particularly in a setting with many covariates. In contrast, GBM is capable of automatically selecting covariates and modeling the treatment selection process without specifying the specific functional form of the model. In this study, there were only 11 covariates in the PS estimation model and two in the ProgPS estimation model. As a result of this low-dimensional setting, GBM's potential may not be fully realized. In other words, GBM may display its advantages to LR in a high-dimensional setting in which it is hard to select covariates and to specify the correct functional forms (e.g., when there are over 100 covariates).

Second, GBM-estimated PSs or ProgPSs might not exhibit their best performance in studies with a small sample. As documented in the PS literature, the PS adjustment is essentially a large sample method (Rubin, 1997). Compared to LR, GBM (and several other machine learning methods) resulted in many more replications that had biased effect estimates at a small sample size (N = 200; Luellen, 2007). Moreover, in simulation studies that compared GBM and LR, the smallest sample size that had acceptable performance for GBM-estimated PSs was N = 368 (Stone & Tang, 2013). In the present study, the specific N varied for each replication, but the largest possible sample size under either of the 3-site conditions was still less than 400, and some replications could even have sample sizes less than 100. In sum, these results suggest researchers may be able to use LR rather than GBM to estimate PSs or ProgPSs when the number of modeling

covariates is relatively few. LR may even be preferred to GBM if the total sample is small. After all, for applied researchers, fitting LR models is technically much less challenging than obtaining the GBM estimates using the *twang* software.

The second factor examined three levels for control group size: 3-site with the contextual factor condition, 3-site without the contextual factor condition, and 38-site condition. The comparison between the 38-site and 3-site conditions allowed for examining the influence of both the total sample size and the ratio between the sizes of the control group and treatment group on the covariate balance and effect estimation. The results revealed that, overall, all methods tended to produce better covariate balance under the 38-site condition than under either of the two 3-site conditions. In terms of effect recovery, the simulation results showed that the impact of the control group size differed between the matching and weighting adjustments. In most cases, matching adjustments produced slightly more accurate estimates under the 38-site condition than under either of the two 3-site conditions. In contrast, weighting adjustments performed much worse under the 38-site condition than under either of the two 3-site conditions.

A tentative conclusion from these results is that a larger total sample size and/or a larger ratio between the control and treatment group sample sizes (as represented by the 38-site condition) is not necessarily beneficial for weighting adjustments based on PSs or ProgPSs. On the other hand, these conditions are preferable for both two-score and one-score matching adjustments. Bai (2015) reported similar findings for the PS-based matching methods and even found that the control group sample size is more influential than the total sample size in terms of bias reduction. However, in this study, both the total sample size and the ratio of control group to treatment group size were larger under the 38-site condition than under either of the two 3-site conditions. As a result, the effect of the control-treatment size ratio from the effect of the total

sample size could not be isolated. Future work could thus examine the impact of sample size ratio with a fixed total sample size, or vice versa.

Furthermore, comparing the two 3-site conditions allowed for examining the impact of the contextual factor on method performance. It is possible that sites may have been sampled disproportionately more in the with contextual factor condition than in the without contextual factor condition. However, the results for all the adjustment methods revealed little difference between the two 3-site conditions in terms of the covariate balance. The difference in effect recovery between these conditions was small as well. These similarities may suggest that similar sites were sampled from the comparison group population in two 3-site conditions.

However, an interesting finding was observed regarding effect estimation. Across all matching adjustments, for the academic outcome, effect size estimates with somewhat lower RMSD were consistently obtained in the condition with the contextual factor. The results were opposite for the disciplinary outcome, for which the effect size estimates had somewhat lower RMSD in the condition without the contextual factor. This suggests that the contextual factor (i.e., the number of years running the program) may be more predictive of the academic outcome than of the disciplinary outcome.

One important implication is derived from these results: similar to local matching, the actual performance of matching within the same social context depends on whether appropriate comparison groups can be found. When researchers seek to construct matches using units from the same social context, they expect this contextual factor (a potential matching covariate) will help create closer matches on the observed and unobserved covariates that are related to the outcomes. However, if the selected contextual factor is weakly related or unrelated to the outcomes, the restriction in the control group pool due to the introduction of the contextual factor will actually

reduce the number of good matches for the treated units. As a result, matching adjustment will perform worse in the condition with the contextual factor than in the condition without it.

Lastly, the sensitivity of the methods to two outcome domains was evaluated. Overall, slightly more biased effect size estimates were found for the disciplinary outcome than for the academic outcome, even though the relative performance of the methods was similar for the two outcome measures. Griffen and Todd (2017) reported similar findings: the PS adjustment methods yielded uniformly larger biases for one outcome measure than for the other. These findings suggest that the effect estimation performance of adjustment methods may also affected by the specific outcome measures of interest.

5.2 LIMITATIONS AND FUTURE DIRECTIONS

As in any study, this study has limitations, and a few of them deserve attention. First, only continuous outcomes were used in the simulations. Many educational researchers are also interested in dichotomous outcomes (e.g., students' promotion or graduation status). Therefore, future research is needed on the performance of two-score methods with dichotomous outcomes. Second, I did not impose a caliper restriction for the matching methods, as I wanted to obtain the same effect estimand across all methods. As discussed earlier, the results for the matching on the ProgPSs may differ if a caliper restriction is introduced. Thus, future research could examine the performance of matching on ProgPSs within a PS or a PROG caliper and could investigate the optimal widths for the PS or PROG caliper for matching on ProgPSs. Third, in this study, the two-score methods were compared to the one-score methods for ATT estimation within a single-level

context. Given that clustered data are very common in educational research, it may also be useful to evaluate the two-score methods' utility within a multilevel framework. To the best of my knowledge, the joint use of PROG and PS with multilevel data was only demonstrated in a simple case study (Kelcey & Swoboda, 2015). Thus, more evaluation work is needed. Lastly, although incorporating one real dataset into the simulation study adds a level of realism to the study, there was a limitation to the number of the pretreatment covariates available for the PS and PROG estimation. Furthermore, the relationships between covariates were fixed and the accuracy of the summary score estimation models employed was unknown. The covariate selection is of critical importance to satisfying the ignorability assumption when summary scores are used to adjust for selection bias. Thus, a comprehensive study of the impact of covariate selection (e.g., omitting the covariates with differential prognostic values), especially in a high-dimensional setting, could contribute to understanding how these methods can best be implemented in practice.

In addition to these study limitations, the knowledge of the joint use of PS and PROG for confounding control is still limited. For example, researchers still know little about the empirical conditions under which the two-score methods yield reliable effect estimates. Scholars have found that Hansen's two-score methods exhibit some advantages over single score adjustments (Hansen, 2006; Leacy & Stuart, 2014), but this advantage does not hold in this study. Thus, more applied and simulation research is needed on the joint use of PSs and PROGs under different conditions to fully explore the two-score method's potential. Other future research could examine different strategies for combining two scores and examine the best conditions for applying joint adjustment methods. In addition, many unanswered questions about PROGs warrant further investigations, including the consequences of overfitting on the PROG estimates, the diagnosis of the misspecified PROG model, and the dimensionality of PROGs for continuous outcomes.

5.3 CONCLUSION

In conclusion, in terms of bias reduction, this study's results do not provide evidence to support the use of any of the examined two-score methods (1:1MAHAL.PS.PROG, FULL.MAHAL.PS.PROG, 1:1M.PROG, and W.ProgPS) as alternatives to adjustments using the PSs or PROGs alone. The adjustment based on PROGs alone (1:1M.PROG) showed some advantages over all the two-score methods and the adjustments using PSs alone (1:1M.PS and W.PS), but its application was limited by the inherent "in-sample" PROG estimation problems and by the lack of reliable diagnostic measures for assessing the validity of the PROG estimates.

One contribution of this study was to expand the understanding of the two-score method, as the results confirm that many features that apply to PSs also apply to ProgPSs (some even applied to PROGs). First, the weighting estimators of both PSs and ProgPSs (W.PS and W.ProgPS) can be very biased and inefficient due to the extreme weights. Second, it may be acceptable to use LR rather than GBM to estimate PSs and ProgPSs when there are few modeling covariates and/or when the total sample size is small. Third, like the matching on PSs, matching on PROGs or ProgPSs works better at larger sample sizes, at higher control-treatment sample size ratios, and with contextual factors related to the outcome when treatment and control units from the same social context are desired. Lastly, this study's results confirm that the pretest measure should be favored when constructing the PS (including ProgPS) and PROG models. For matching on the PSs or PROGs only, or for the joint use of these scores through matching, greater balance on the pretest measures indicates more accurate effect estimates.

In addition, this study's findings also lead to one suggestion that is inconsistent with the common practices for the PS adjustment. Specifically, when selecting the model specification for

a summary score and/or selecting the summary score's estimation-adjustment combination, researchers should not choose the methods that produce the best average balance across all measured covariates or those that produce the fewest measured covariates with large differences. Instead, it may be better to choose the procedures that maximize the degree of balance in the pretest measures and in other prognostically important covariates.

Overall, despite the lack of promising results for the two-score method, more research is necessary before concluding that two-score adjustments are no better than single score adjustments. As in any simulation study, this study's results and inferences are restricted to its specific settings. Thus, any recommendations are limited to this context, and more research is needed to generalize the findings.

BIBLIOGRAPHY

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics, 86*(1), 180-194.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation efficacy of a university-level remedial writing program. *Evaluation Revie*w, *22*(2), 207-244.

An, B. P. (2012). The impact of dual enrollment on college degree attainment: Do low-SES students benefit? *Educational Evaluation and Policy Analysis*, *35*(1), 57-75.

Arbogast, P. G., & Ray, W. A. (2009). Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research, 18*, 67-80.

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics - Simulation and Computation, 38*(6), 1228-1234.

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. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behavioral Research, 47*(1), 115-135.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26*(4), 734-753.

Austin, P. C. & Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research*, *26*(4), 1654–1670.

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

Bernstein, L., Dun Rappaport, C., Olsho, L., Hunt, D., & Levin, M. (2009). *Impact evaluation of the U.S. Department of Education's student mentoring program* (NCEE 2009-4047). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management, 31*(3), 729-751.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.

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

Cadarette, S. M., Gagne, J. J., Solomon, D. H., Katz, J. N., & Stürmer, T. (2010). Confounder summary scores when comparing the effects of multiple drug exposures. *Pharmacoepidemiology and Drug Safety, 19*(1), 2-9.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72.

Cham, H. (2013). *Propensity score estimation with random forests* (Unpublished doctoral dissertation). Arizona State University, Tempe.

Cham, H. & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods, 21*(3), 427-445.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295-313.

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

Cole, S. R. & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology, 168*(6), 656-664.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.

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

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: the relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods, 15*(1), 56-68.

D'Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265-2281.

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

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49*(4), 1231-1236.

Dong, D., & Lipsey, M. (2014). *How well propensity score methods approximate experiments using pretest and demographic Information in educational research*? Paper presented at the 35th Annual Association for Public Policy Analysis and Management (APPAM) Research Conference, Albuquerque, NM. Retrieved from http://files.eric.ed.gov/fulltext/ED562732.pdf

Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). *Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates* (NCEE Technical Methods Report No.2012-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garrett, R., & Hong, G. (2016). Impacts of grouping and time on the math learning of language minority kindergartners. *Educational Evaluation and Policy Analysis, 38*(2), 222-244.

Glynn, R. J., Gagne, J. J., & Schneeweiss, S. (2012). Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and Drug Safety, 21*(S2), 138-147.

Griffen, A. S. & Todd, P. E. (2017). Assessing the performance of nonexperimental estimators for evaluating Head Start. *Journal of Labor Economics, 35*(1), S7-S63.

Gu, X., & 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*. Thousand Oaks, CA: Sage.

Hallberg, K. (2013). *Identifying conditions that support causal inference in observational studies in education: Empirical evidence from within study comparisons* (Unpublished doctoral dissertation). Northwestern University, Evanston.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*(467), 609-618.

Hansen, B. B. (2006). *Bias reduction in observational studies via prognostic scores*. (University of Michigan, Statistics Department, Technical Report 441). Retrieved from http://www.stat.lsa.umich.edu/~bbh/rspaper2006-06.pdf

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika, 95*(2), 481-488.

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.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*, 107-128.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hernandez, J. M. (2015). *Causal inference using educational observational data: Statistical bias reduction methods and multilevel data extensions* (Unpublished doctoral dissertation). University of Washington, Seattle.

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. W., & 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 Analysis, 15*(3), 199-236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

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 of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901-910.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of Royal Statistical Society, Series A (Statistics in Society), 171*(2), 481-502.

Kelcey, B. M. & Swoboda, C. M. (2015). Prognostic scores in clustered settings. In W Pan & H Bai (Ed.), *Propensity score analysis: Fundamentals and developments* (pp. 348-376). New York, NY: The Guilford Press.

Keller, B. (2013). *Data mining alternatives to logistic regression for propensity score estimation* (Unpublished doctoral dissertation). University of Wisconsin-Madison, Madison.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *The American Economic Review, 76*(4), 604-620.

Leacy, F. P., & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine, 33*(20), 3488-3508.

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

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

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

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 the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine, 23*(19), 2937-2960.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403-425.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of Education, 74*(4), 341-374.

Normand, S. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiograph following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*, 387-398.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis, 36*(2), 127-144.

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology, 181*(2), 108-119.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal Inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*(4), 463-479.

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *31*, 172-181.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2017). *Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package*. Retrieved from https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

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

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine, 115*(11), 901-905.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-52.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*(366a), 318-328.

Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. *Biometrics*, *36*(2), 293-298.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which 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*, 757-763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100*(469), 322-331.

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*(1), 20-36.

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. (2017). *Base SAS® 9.4 procedures guide* (7th ed.). Cary, NC: SAS Institute Inc.

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

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*(6), 546-555.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology, 60*(1), 607-629.

Shadish, W. R., Clark, M., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1344.

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

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation, 35*(3), 311-327.

Steiner, P. M., & Cook, T. D. (2013). Matching and propensity scores. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Foundation* (Vol. 1, pp. 237–259). New York, NY: Oxford University Press.

Steiner, P. M., Cook, T. D., Li, W., & Clark, M. H. (2015). Bias reduction in quasi-experiments with little selection theory but many covariates. *Journal of Research on Educational Effectiveness, 8*(4), 552-576.

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., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250-267.

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

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

Stuart, E. A. & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics***,** *33*(3), 279-306.

Stürmer, T., Schneeweiss, S., Brookhart, M. A., Rothman, K. J., Avorn, J., & Glynn, R. J. (2005). Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology, 161*(9), 891-898.

Sullivan, A. L., & Field, S. (2013). Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology, 51*(2), 243-260.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118.

Tu, C., & Koh, W. Y. (2017). A comparison of balancing scores for estimating rate ratios of count data in observational studies. *Communications in Statistics - Simulation and Computation*. *46*(1), 772-778.

Wang, X. (2014). Pathway to a baccalaureate in stem fields: Are community colleges a viable route and does early STEM momentum matter? *Educational Evaluation and Policy Analysis, 37*(3), 376-393.

Watkins, S., Jonsson-Funk, M., Brookhart, M. A., Rosenberg, S. A., O'Shea, T. M., & Daniels, J. (2013). An empirical comparison of tree-based methods for propensity score estimation. *Health Services Research, 48*(5), 1798-1817.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology, 63*(8), 826-833.

What Works Clearinghouse (2017). *Procedures handbook* (*version 4.0*). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management, 26*(3), 455-477.

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness, 10*(1), 207-236.

Wyss, R., Glynn, R. J., & Gagne, J. J. (2016). A review of disease risk scores and their application in pharmacoepidemiology. *Current Epidemiology Reports, 3*(4), 277-284.

Wyss, R., Hansen, B.B., Ellis, A. R., Gagne, J. J., Desai, R. J., Glynn, R. J., & Stürmer, T. (2017). The "dry-run" analysis: A method for evaluating risk scores for confounding control. *American Journal of Epidemiology, 185*(9), 842-852.

Xu, D., & Jaggars, S. S. (2011). The effectiveness of distance education across Virginia's community colleges: Evidence from introductory college-level math and English courses. *Educational Evaluation and Policy Analysis, 33*(3), 360-377.