ABSTRACT

Title of Dissertation:	THE EFFECT OF DIFFERENT RELATIVE LOGISTIC REGRESSION GENERATED PROPENSITY SCORE DISTRIBUTIONS ON THE PERFORMANCE OF PROPENSITY SCORE METHODS
	Ji An, Doctor of Philosophy, 2020
Dissertation directed by:	Laura M. Stapleton, Professor Department of Human Development and Quantitative Methodology, Measurement, Statistics and Evaluation Program

Much education research involves evaluating the causal effects of interventions. The propensity score (PS) method, which is often used to account for selection bias, has become a popular approach to facilitating causal inference in quasi-experimental designs. Because the success of the application of PS conditioning methods is dependent on the estimated propensity scores, the relative PS distribution between the treated and control groups could be an important yet not well-known factor. The primary goal of this dissertation was to explore, via a simulation study, the relations between the relative PS distributions and the performance of selected PS matching methods. The results indicated that PS weighting (without trimming) tends to be robust to a variety of data conditions and produces more accurate and trustworthy TE and SE estimates. The performance of the methods and conclusions were then illustrated through an empirical data analysis using data selected from the Early Childhood Longitudinal Study Kindergarten Class of 2010-11 study, assessing the effect of having home computers on first grade students' math achievement.

THE EFFECT OF DIFFERENT RELATIVE LOGISTIC REGRESSION GENERATED PROPENSITY SCORE DISTRIBUTIONS ON THE PERFORMANCE OF PROPENSITY SCORE METHODS

by

Ji An

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Laura M. Stapleton, Chair
Professor Gregory R. Hancock, Co-chair
Professor Jeffrey Harring
Professor Hong Jiao
Professor Yan Li, Dean's Representative

DEDICATION

This dissertation is dedicated to my mother, my husband, and my daughter.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Laura M. Stapleton for her continuous support of my Ph.D. study and research. What I learned from her was far beyond what was written in this dissertation. I have been used to looking for encouragement, confirmation, and praise in your eyes whenever I feel unsettled, to gain courage. I will always remember your smiley look when I tell something as a good story, pursue a higher step in my career, and handle tough moments in my life with a light heart. Laura and Ji will continue to rock the world together, because family never leaves you.

I would also like to extend my deepest gratitude to Dr. Gregory R. Hancock. I feel greatly honored to have the opportunity of working with you closely in the past few years. You are like a magician who is able to turn anything bland into magic, with a sense of humor. This wonderful experience has been written in our "HA method", as well as many other research papers. I am so happy.

I would like to express my sincere appreciation to the other professors on my committee, Dr. Jeffrey Harring, Dr. Hong Jiao, and Dr. Yan Li. The completion of my dissertation would not have been possible without your support and advice.

I came across countless smart people since I started my career in industry, as smart as my EDMS professors. In a bigger world, talents can be easily found, however, my professors have always been spectacular and irreplaceable with their generocity in love and protection, their sense of scientific rigor, and many other latent features that made them spectacular and irreplaceable. Whenever I accomplish something that I'm proud of, I could always tell who I learned that from. Thank you, EDMS. I am indeed grateful for this precious journey.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1. Introduction

Much education research involves evaluating the effect of interventions, such as the effect of a retention program or the National School Lunch Program, on student achievement. The ideal design for making causal inference regarding such topics is the experimental design where all subjects are randomly assigned to the treated and control groups, which, unfortunately, is not always feasible. For example, we could neither ethically randomly hold back students nor randomly assign students to receive free or reduced-price lunch without considering the household income level. In such quasi-experimental designs, in which it is not feasible to control the assignment of the experimentation and units are not allocated randomly to the two groups, the treatment group assignment tends to depend on the pretreatment characteristics of group members; thus differences between groups can be attributed in part or possibly entirely to the pretreatment characteristics as opposed to the treatment itself (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007).

According to the *What Works Clearinghouse Standards Handbook* (2017), only when the confounding effects in quasi-experimental designs are appropriately considered can the causal inference of the related research meet group design standards with reservations. This goal, as indicated in the *Handbook*, can be achieved mainly by controlling for the confounding factors via obtaining baseline equivalence between groups. An approach that has received a great surge of interest in facilitating valid causal inference in quasi-experimental designs is the propensity score (PS) method (Stuart, 2010).

The PS method, a widely-used approach in social sciences and education to facilitate causal inference, can effectively account for selection bias in quasi-experimental settings. In particular, a typical PS procedure involves the selection of the confounding pretreatment covariates that are believed to affect the nonrandom assignment to the treatment group in a quasi-experimental design. The probability of being assigned to the treatment condition for each person (i.e., the PS estimate), then, can be estimated as a function of the selected covariates. Following this PS estimation process, a PS conditioning method is conducted to make sure that only those with similar PS estimates (and thus similar probabilities of receiving the treatment) are compared between the treatment and control conditions. The treatment effect can then be estimated based on the conditioned data, in which the selection bias has been controlled.

According to the above brief description of the PS method, PS conditioning is the key step to minimize the selection bias so as to facilitate causal inference. The fact that the PS conditioning process is highly dependent on the PS estimates makes the relative PS distribution between the treated and control groups an important, yet not well known, factor in the generation of appropriate causal estimates. The primary objective of this dissertation is to further explore the use of the PS methods and to disseminate the findings of the effects of different relative PS distributions on causal estimates to provide guidance to education researchers on PS application. This will include dissemination on what type of PS conditioning methods you should use depending on the PS distributions and how the heterogeneity of treatment effect (TE) impacts the performance of PS methods. Findings from this study will add to the literature on the implementation of the PS method – one more step of relative PS distribution checking and PS conditioning

method selection might be necessary as opposed to the traditional PS procedure where relative PS distributions is never a concern.

In addition to the methodological perspective taken in this dissertation, an empirical study will demonstrate the steps and implementation considering PS distributions. The goal is to provide a guide, or example, to other researchers who are using national data for causal effect analyses that utilize PS methods. The results could also serve as a resource for future educational technology related policy making (e.g., broadening home computer access to first grade students in the United States to help their academic achievement).

This dissertation starts with an overview of the PS methods, including the potential outcomes framework and key assumptions, as well as the implementation procedure of different PS methods, followed by the importance of relative PS distributions on TE estimation. Chapter 3 describes the detailed research design of the simulation study and discusses the limitations of the simulation design, followed by interpretation of results in Chapter 4. Chapter 5 introduces the data source and design of the empirical demonstration, as well as illustrates the results and inferences. Finally, Chapter 6 summarizes the results, limitations of the study, and talks about the possibilities for future research.

Chapter 2. Conceptual Framework and Literature Review

As previously mentioned, the main objective for the application of PS methods is to identify the causal relations between the treatment and the outcome in quasi-experimental designs. The first part of this chapter briefly introduces the potential outcomes framework for estimating causal inference. I then provide an overview of the PS methods, including the key concepts, recent developments, and assumptions, and then I elaborate on each step of the implementation procedure. I continue the chapter with a detailed description of the potential implications of PS relative distributions and heterogeneous treatment effects that motivated this study. Finally, I conclude the chapter with the two research questions that are addressed in this study.

2.1 Causal Inference and the Potential Outcomes Framework

In the potential outcomes framework, two treatments are available, the active treatment (i.e., treatment) and the control treatment (i.e., control). Each individual thus potentially has two outcomes, the outcome under the treatment and that under the control. First proposed by Rubin (1974), a causal effect can be calculated by comparing the two potential outcomes. In particular, the causal effect of the treatment on unit i is simply the difference between i's outcomes from receiving treatment $Y_i(1)$ and receiving control $Y_i(0)$, that is,

$$\delta_i = Y_i(1) - Y_i(0). \quad (1)$$

Unfortunately, each individual can receive either treatment or control and thus only one outcome can be observed (Holland, 1986). For example, for a subject assigned to the treatment condition, only $Y_i(1)$ is available and the counterfactual outcome $Y_i(0)$ only exists under the unobserved condition. Therefore, the individual causal effect of treatment cannot be determined. When units are randomly assigned to treatment or control conditions, however, the subjects in

one group can be considered the counterparts of those in the other group, and thereby, one can still estimate the treatment effect via comparing the two groups. The average causal effect, in turn, is simply the difference between the outcomes of the two groups, as expressed in the following, under the assumption that the two groups are "equal in expectation" (Murnane & Willett, 2010, p. 31):

$$\delta = E\left[Y(1)\right] - E\left[Y(0)\right]. \tag{2}$$

However, in a non-experimental design, groups may differ systematically with respect to unobserved pretreatment covariates and are incomparable. Thus, it is desirable to replicate an experimental design as closely as possible via obtaining treated and control groups with similar distributions on each pretreatment covariate (Stuart, 2010), and this goal can be achieved by using PS methods.

2.2 An Overview of the Propensity Score Methods

The following sections introduce the recent developments of the PS methods. The major assumptions, as well as the specific implementation procedures are also described in detail.

2.2.1 Propensity score methods and recent developments

A propensity score, defined as the conditional probability of a participant to be assigned to the treatment condition (Rosenbaum & Rubin, 1983), is considered a composite of all pretreatment variables that describes the selection process (Murnane & Willett, 2011). In particular, conditional on the propensity scores obtained from the correct PS model, the selection bias is controlled, so that the distributions of the baseline covariates between the control and treatment groups are similar as in a randomized design (Austin, 2011a). The assumption in using this method are further described in Section 2.2.2.

Alternatives to the PS methods include direct control for covariates using regression analysis, instrumental-variables estimation, structural equation modeling, and so on. The PS methods have shown advantageous features compared to the other methods. First, taking the most popular multiple regression approach as an example, it can be problematic when the number of covariates is large. Imagine that in a study where there are 30 covariates, if all these covariates and their two-, three-, four-, and multi-way interactions are included as controls in a regression analysis, it may eventually run out of degrees of freedom, which may in turn overestimate the standard errors and dramatically reduce the power (Murnane & Willett, 2011). Compared to regression, the PS methods focus on a balancing score, that is, a combination of the covariates rather than controlling for all related covariates. In other words, the purpose of estimating propensity scores is to obtain balance between the two groups as opposed to estimating precisely and using any parameter coefficients. This idea of a balancing score is to a large degree robust to the overfitting issue, as the precision and the power in the PS estimation step are not concerns. A second concern with the use of regression approaches, is that sufficient overlap of the covariates between the two groups have been shown an important factor for obtaining accurate TE estimates (Dehejia & Wahba, 1999; Glazerman, Levy & Myers, 2003). In the TE estimation process, regression-based approaches may be highly dependent on model extrapolation, which can give rise to biased estimates (Arpino & Cannas, 2015; Drake, 1993). This is the case especially when the two treatment groups have substantial differences, in other words, insufficient overlap (also referred to as a lack of common support), in terms of the covariates. This issue of the regression-based approaches is also not accounted appropriately in the alternative methods. The PS methods, on the other hand, involve a diagnostic procedure of

checking the common support, which helps reduce unrealistic extrapolation and improves estimation accuracy.

In recent years, developments and applications of PS methods have occurred in many scientific areas including medical research for evaluation of medical outcomes (Stone, Obrosky, Singer, Kappor, & Fine, 1995), empirical labor economics for evaluation of labor market policies (e.g., Bryson, Dorsett, & Purdon, 2002; Dehejia & Wahba, 1999; Heckman, Ichimura, & Todd, 1998), educational and psychological research to study the effect of programs on students' performances (e.g., Hong & Raudenbush, 2006; Morgan, Frisco, Farkas, & Hibel, 2010), and so on. In addition to evaluation of interventions, use of the propensity score methods has also increased in psychometrics for the assessment of students' ability (e.g., Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Guo, 2009; Jiao, Zou, Liao, Li, & Lissitz, 2016; Li, Liao, Zou, Jiao, & Lissitz, 2016; Sireci, 1997).

Given that PS methods are experiencing such a tremendous increase of interest, there are several published instructive works that focus on implementation guidance and suggestions for the PS analysis (Austin, 2011a; Caliendo & Kopeinig, 2008; Harder & Stuart, 2010; Stuart, 2010; Shadish & Steiner, 2010). At the same time, a variety of specific practical issues have arisen from PS applications that have attracted PS methodological researchers' attention and lead to research on many specific topics. Examples include PS model specification (Drake, 1993; McCaffrey, Ridgeway, & Morral, 2004), caliper selection for PS matching (Austin, 2011b; Cochran & Rubin, 1973; Rosenbaum & Rubin, 1984), weight trimming strategies for PS weighting (Lee, Lessler, & Stuart, 2011), the application of PS methods using data that come from a complex sampling structure (DuGoff, Schuler, & Stuart, 2014; Hahs-Vaughn, 2015; Lee, Lessler, & Stuart, 2010; Thoemmes & West 2011), and Bayesian PS estimation (Alvarez &

Levin, 2014; Kaplan & Chen, 2012; McCandless, Gustafson, & Levy, 2009a; McCandles et al., 2009b). Although a broad array of topics have been studied, there are still remaining questions yet to explore, including the interest of this study – the use of PS methods with different relative PS distributions. This issue will be further discussed in Section 2.3.

2.2.2 Assumptions

A key assumption of PS methods related to PS model estimation is *strong ignorability*, that is, all nonequivalence between the treated and control groups (except for the nonequivalence in the outcome measures) is removed after controlling for the pretreatment covariates. The assumption of *ignorability* has two components: (a) the potential outcomes are independent of the treatment status conditioning on the pretreatment covariates and (b) every unit has a positive (i.e., nonzero) probability of receiving the treatment. These ideas can be expressed as $[Y_i(1), Y_i(0)] \perp T_i|X_i$ and $0 < P_i(T_i = 1|X_i) < 1$. These two components together imply that the PS analysis will yield unbiased estimates only if all confounding covariates (including the interactions among the covariates), observed and unobserved, are controlled and there is at least some overlap between the two treatment groups (Rosenbaum & Rubin, 1983).

Because neither the covariates used in treatment selection nor the relations among the covariates are known (Drake, 1993), two steps are required to meet this assumption and successfully obtain the PS estimates: identifying all possible pretreatment covariates and specifying the accurate functional form of the PS models. Sensitivity analyses (Imbens, 2004; Rosenbaum & Rubin, 1983; Rosenbaum, 2002) are available to examine the extent to which the inference about the causal effect would change if there were unobserved confounders. As will be addressed later, the use of sensitivity analysis was not part of this dissertation.

Another assumption to obtain unbiased estimates of causal effects using PS methods is referred to as the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980). To be more specific, it requires that the assignment of an individual does not affect the outcome of any other individuals. SUTVA can be violated due to "spill over" effects, such as siblings assigned to different treatment groups talk about it at home and affect each other's outcome. Another example of violation of SUTVA is that a student's math score is improved only because his/her friend joined the same math tutoring program and they often help each other. Rubin also pointed out that SUTVA is difficult to claim in educational (e.g. school and classroom) type settings because children always interact with each other. Regardless of the plausibility of SUTVA in educational settings, the chance of violation of this assumption can be decreased by taking careful control over the experiment and reducing the interaction between the two treatment groups (Stuart, 2010).

2.2.3 Propensity score implementation (5 steps)

As was briefly introduced in Chapter 1, a typical PS procedure involves five steps, which include (1) identifying appropriate pretreatment covariates that could potentially cause bias in the TE estimate, if ignored; (2) estimating propensity scores via, typically, a logistic regression on the baseline covariates; (3) conditioning on (or matching on) propensity scores between the treated and control groups; (4) assessing the conditioning quality (i.e., balance check); and (5) estimating the treatment effect based on the conditioned, or matched, sample (Stuart, 2010; Thoemmes & West, 2011). Each of these steps is described in the sections that follow.

2.2.3.1 Identification of critical covariates. Identifying appropriate covariates that could potentially cause bias in the TE estimate is the first step of a PS analysis. The key point for an effective PS analysis is to select appropriate covariates. Stuart and Rubin (2007) suggested two

points that need to be carefully considered when selecting covariates. First, the covariates selected should be able to yield desired good matches between the treatment and control groups. In particular, the covariates selected must be related to the treatment assignment and it is important to include a large set of covariates (Hill, Reiter, & Zanutto, 2004; Rubin & Thomas, 1996). Hill, Reiter, and Zanutto (2004) and Rubin and Thomas (1996) suggest to include as many covariates as possible. Greevy, Lu, Silber, and Rosenbaum (2004) concluded from an empirical study that including a larger set of covariates yields greater power, even when some covariates are indirectly related to the treatment through their correlation with the other directly related variables. Second, the covariates selected should not be affected by treatment assignment. This is because the propensity score represents the probability of treatment exposure conditional on covariates (Lunceford & Davidian, 2004). Treatment must happen after the covariates. Otherwise, matching can lead to substantial bias in the estimated treatment effect (Frangakis & Rubin, 2002; Imbens, 2004; Stuart & Rubin, 2007).

2.2.3.2 Estimation of the PS. As was introduced earlier, the traditional way of obtaining propensity scores is to fit a logistic regression on all pretreatment covariates. The estimated propensity score is then the predicted probability of being in the treatment. The overall PS model can be expressed as:

$$logit\left(e_i\right) = \beta_0 + \sum_{p=1}^{P} \beta_p X_{pi}, \tag{3}$$

where e_i is the propensity score for person i, β_0 is the intercept of the model, β_p represents the regression coefficients for the p^{th} covariate, and X_{pi} is the score of the p^{th} covariate (or potential interaction among or polynomials for particular covariates) for subject i.

In the case where more than two treatment groups exist, a few options are available for the PS estimation, such as to estimate the multinomial logit, the multinomial probit, or a series of

binomial models (Lechner, 2001). The advantages and disadvantages of these options are not further discussed here as this study only focused on the situation where there are only two treatment groups.

The PS estimation method based on logistic regression can be challenging in practice. As was discussed earlier, for the PS method, the assumption of ignorability implies that making an unbiased inference via PS analysis not only requires one to find all the related covariates, but also suggests to appropriately identify the nonlinear components such as polynomial and interaction terms. Since the true PS model is almost always unavailable, researchers have tried the following methods to approach it.

The most commonly considered criterion for nonlinear term examination is to directly check the balance between the two treatment groups after PS adjustment. The idea originally articulated by Rosenbaum and Rubin (1984) is a stepwise procedure. In general, it involves fitting a PS model with main effects only based on the selected covariates and stratifying the propensity scores into five subclasses testing for covariate differences between the two groups within each subclass (i.e., F test for group comparison in terms of variance; t test can be another option for checking group mean differences), and adding polynomial or interaction terms to the previous PS model where significant group differences still exist. This procedure is iterated until there is no statistically significant group differences. Similar methods have been implemented in Dehejia and Wahba (1999) and Mojtabai and Graff Zivin (2003).

Another nonlinear term selection procedure proposed by Hirano and Imbens (2001) does not involve checking the covariate balance. Instead, it suggests to lay out all possible predictors of the propensity score model, including the main effects of the pretreatment covariates, higher order terms of the covariates, possible interactions of the covariates, and selected higher order

terms of the interactions. The next step is to run logistic regressions with a single predictor from all the above candidate predictors one by one and keep the ones with t statistics larger in absolute values than some pre-specified cutoff value. Despite decreasing the possibility of model misspecification to the most extent, the authors admitted that this method can be prohibitively expensive, especially when the number of covariates is relatively large.

The above parametric logistic regression methods with selected polynomials and interactions have been used almost exclusively in existing literature. However, even if these proposed methods are applicable, most applications still assume that the PS model follows an additive and linear logistic regression on the log-odds scale, probably due in large part to its complexity (McCaffrey et al., 2004). In 2004, McCaffrey and colleagues started to comprehensively explore the possibility of using data mining techniques (also referred to as nonparametric methods; e.g., regression trees, random forest, boosted regression, etc.) for PS estimation without considering model specification, followed by Setoguchi, Schneeweiss, Brookhart, Glynn, and Cook (2008); Lee et al. (2010), and Cham (2013). These methods do not assume any functional form between the covariates and the binary treatment outcome yet are able to capture complex relations among the covariates.

This simulation study only focused on the PS estimation based on logistic regressions. Because the true PS model was known, on the one hand, implementing an accurate PS model using logistic regression would isolate the noise caused by PS misspecification from the results. On the other hand, a PS misspecification scenario was also simulated in the study to investigate how it would interact with the relative PS distributions and affect the accuracy and precision in the final inference.

2.2.3.3 PS conditioning. Once propensity scores are obtained, the PS *conditioning* procedure (also referred to as *matching* in general in the literature) is implemented to make the treatment and control groups comparable. This is the key step to obtain group equivalence on each pretreatment covariate, which in turn allows for the two groups to be considered comparable as in an experimental design. A variety of PS conditioning techniques are available in practice and the most popular ones include PS matching (the *PS matching* here refers to a specific technique of PS conditioning), subclassification, and weighting.

PS matching. The goal of PS matching is to obtain a sample in which the subjects from the treated and control groups are similar on their propensity scores, with the unmatched cases discarded (Rosenbaum & Rubin, 1985; Rubin, 1973). Thus, the PS matching method does not necessarily use all of the data.

There are a variety of options for the implementation of PS matching, with the most important decisions being the following: the measures of similarity (or "distance") between propensity scores (i.e., exact matching to match directly on the PS estimates and matching on the transformed propensity scores such as Mahalonobis matching), the size of the matched samples (e.g., 1:1 or 1:k matching, where $k = 2, 3, \dots, K$, depending on the ratio of the target matched group sizes), whether to restrict the distances between the matched individuals (i.e., caliper matching), whether to use the sample repeatedly (i.e., matching with or without replacement), and which matching algorithm to use (e.g., greedy/nearest neighbor matching or optimal matching depending on how the researcher wants to minimize the differences between the matched samples). For example, the 1:1 nearest neighbor caliper matching without replacement matches each subject in the treatment group with the smallest distance (nearest neighbor) from an individual in the control group; a treatment observation with a distance to the nearest neighbor

that is larger than the pre-specified caliper (e.g., 0.25) is discarded; the process is done without replacement, that is, an untreated subject that has been matched to a treated subject is no longer eligible to be matched to other treated subjects.

Austin (2014) provided a comprehensive overview of the matching method and examines the performance of different combinations of the PS matching options introduced above via Monte Carlo simulations and provides suggestions for practice. One of the important findings from Austin was that greedy nearest neighbor caliper matching without replacement (with subjects chosen for matching in a random order or the order of the best possible match) tended to yield TE estimates with minimal bias across a wide range of scenarios.

In addition to the regular matching methods introduced above, Iacus, King, and Porro (2012) recommended a new matching method known as "Coarsened Exact Matching" (CEM) that belongs to a new generalized class of matching methods known as "Monotonic Imbalance Bounding" (MIB). Iacus et al. argued that the traditional matching methods depend on an ex ante process where the level of reduction in imbalance between the two groups is not guaranteed for all pretreatment covariates. As a result, many applications of matching methods that fail the balance check need to be repeatedly tweaked and rerun, sometimes even making the balance on some of the covariates worse and so could be the accuracy of the TE estimates. With CEM, rather than matching all covariates all together based on the PS estimates, each covariate is coarsened by recoding depending on the practical meaning of the covariate, so that values are grouped and assigned the same value (e.g., years of education are recoded into high school, college, and post-graduate degrees); then perform the exact matching algorithm to prune the unmatched data. That said, the improvement on balance of one covariate does not affect the balance of the other covariates. Although CEM has been demonstrated more efficient and can

produce better balance in many cases, it was not considered in this study, as choosing the appropriate coarsening is challenging without knowing the practical meaning of the covariates with simulated data. In addition to matching, another popular PS conditioning method is subclassification.

Subclassification. Subclassification, also known as stratification, involves stratifying all subjects into multiple mutually exclusive subsets based on percentiles of the PS estimates (Austin, 2011a; Rosenbaum & Rubin, 1984). The treated and control subjects should then have similar and thus comparable propensity scores within each subclass. Once the subclasses are created, a subclass-specific treatment effect can be estimated within each subclass by directly comparing the means of the outcome measures between the two groups. Each subclass-specific TE estimate can then be pooled to produce an overall treatment effect. Pooling can be achieved by weighting each estimate by the inverse the proportion of subjects that are classified into that subclass. Two kinds of weighting are available in the PS literature depending on which population is of interest: all individuals or just those who are treated (the distinction between these two populations is discussed in Section 2.2.3.5). These two types of subclassifications can be expressed in the following equations:

$$\delta_{all} = \sum_{k} \left(\frac{n_k}{n} \right) \delta_k , \tag{4}$$

$$\delta_{treat} = \frac{1}{k} \sum_{k} \left(\frac{n_{kT}}{n_k} \right) \delta_k , \tag{5}$$

where δ_{all} and δ_{treat} are the two types of treatment effects mentioned above and will be further introduced in Section 2.2.3.5; δ_k is the treatment effect for subclass k; n_k,n_{kT}, and n

respectively represent the sample size for subclass k, the sample size for the treated subjects in subclass k, and the total sample.

It was found by Rosenbaum and Rubin (1984) that creating five subclasses using the quintiles removes more than 90% of the bias in the TE estimates and subclassifying on quintiles has been used by researchers as a rule of thumb. More recent work (e.g., Lunceford & Davidian, 2004; Huppler Hullsiek & Louis, 2002) suggest that stratifying on quintiles may not be ideal with increasing sample size, however, guidelines for choosing the number of subclasses depending on sample size have not been established. Therefore, this paper still followed the most popular five-subclass recommendation by Rosenbaum and Rubin, utilizing both weighting strategies mentioned above to address the two different population inferences.

PS weighting. PS weighting is similar to the idea of weighting a sample selected following a complex sampling design so that weighted estimates appropriately represent specific population parameters (Morgan & Todd, 2008). Two PS weighting methods, inverse probability of treatment weighting (IPTW) and weighting by the odds (WBO), are available also depending on whether all individuals or just those who are treated are of interest in the study. Again the distinction in these two inference populations is illustrated in detail in Section 2.2.3.5.

Each subject's IPTW weight is the inverse of the probability of receiving the treatment that the individual actually received. Specifically, as expressed in Equation 6, subjects in the treated group are weighted by the inverse of their propensity scores and those in the control group are weighted by the inverse of the probability of not receiving the treatment (Czajka, Hirabayashi, Little, & Rubin, 1992; Imbens, 2004).

$$w_{\text{IPTW},i} = \frac{T_i}{e_i} + \frac{1-T_i}{1-e_i}, \tag{6}$$

where $w_{\text{IPTW}.i}$ is the IPTW weight for the ith subject, T_i is an indicator variable denoting whether or not subject i received the treatment, e_i is the estimated propensity score for subject i.

With WBO, as expressed in Equation 7, the treated subjects receive a weight of 1 while the control subjects are weighted by the odds of being treated (Harder et al., 2010; Hirano et al., 2003). This way, both groups are weighted to the treatment group and thus gives the TE estimate for the population that was selected to the treatment.

$$w_{\text{WBO}.i} = T_i + \frac{(1-T_i)\times e_i}{1-e_i}, \tag{7}$$

where $w_{\text{WBO}.i}$ is the WBO weight for the i^{th} subject and remaining are defined as in Equation 6. The same as subclassification, with PS weighting, all subjects are retained in the analyses. In this simulation study, these PS conditioning methods including 1:1 nearest PS matching, subclassification, as well as PS weighting for both treatment effects were all implemented and compared in terms TE accuracy and precision.

In summary, in the PS conditioning step, the data from the two treatment groups that have been adjusted for TE estimation via any of the conditioning methods are referred to as conditioned data. The goal of these PS conditioning methods is to achieve balance between the conditioned data such that the treated and control groups are comparable, although a good balance is not guaranteed. Therefore, balance check is another important step before moving forward to estimating unbiased treatment effects.

2.2.3.4 Assessing the conditioning quality. The idea of balance check is to see if the conditioning procedure is able to obtain balance between control and treated groups on the distributions of the relevant covariates (Caliendo & Kopeinig, 2008). A typical procedure is to check the standardized bias of the covariates in the treatment group compared to those in the control group (Rubin, 1985). Standardized bias is often defined as the standardized mean

difference (SMD) between control and treated groups, a quantity similar to a standardized effect size (Cohen's *d*). Covariates a considered balanced if SMD is less than 0.25 (Ho, Imai, King, & Stuart, 2007). Usually, an even stricter cutoff, such as 0.10, indicates better balance (Harder, Stuart, & Anthony, 2010). SMD can be calculated from:

$$SMD_p = \left| \frac{\bar{X}_{pT} - \bar{X}_{pC}}{S_{pT}} \right|, \quad (8)$$

where $\bar{X}_{pT}$ and $\bar{X}_{pC}$ are the means of covariate p in the treated and control groups and s_{pT} is the standard deviation of the covariate within the treatment group. For IPTW and WBO, the components are weighted equivalents. For subclassification, the balance within each stratum is calculated; the weighted mean of all balance measures across strata is then computed. If conditioned samples are found to be unbalanced on one or more covariates, a common solution is to reconsider what other pretreatment covariates can be included and/or add possible interactions or polynomials in the PS estimation model. In this simulation study, balance was always checked after PS conditioning. Given that all important pretreatment covariates were included in the PS models, balance was mostly desirable. In the conditions where balance was not good, I moved forward to TE estimation without going back to the PS model again and discussed this in the results.

2.2.3.5 Treatment effect estimation. After covariate selection and PS estimation, conditioning, and balance checking, the last step in applying PS techniques is to obtain and interpret the TE estimate. As was mentioned in Section 2.2.3.4, there are two causal estimands commonly of interest in quasi-experimental settings depending on the population of interest: the average treatment effect on the treated (ATT, Equation 9) and the average treatment effect in the population (ATE, Equation 10).

$$ATT = E\left[Y_i(1) - Y_i(0) | T = 1\right]. \quad (9)$$

$$ATE = E\left[Y_i(1) - Y_i(0)\right]. \quad (10)$$

The ATT is the treatment effect for the subjects in the treated group only (Hirano, Imbens, & Ridder, 2003; Imbens, 2004). The ATE refers to the treatment effect on all subjects, both treated and control (Czajka et al., 1992; Imbens, 2004). In the math tutoring example, ATT compares the math achievement between students who had taken the treatment and the math achievement for the same students if, instead, they did not take the program. ATE, in comparison, represents the difference in math achievement if all students took the program compared to if none of them were in the treatment program. As is seen in Equation 10, the definition of ATE is identical to that of the general causal effect in the potential outcomes framework.

The PS conditioning methods (matching, subclassification, and weighting) differ in terms of the estimands, for the reason that they vary with respect to the relative weights each subject receives as well as the number of subjects that retain after conditioning (Stuart, 2010). In particular, the PS method based on matching produces an estimate of the ATT, while subclassification and weighting can estimate both the ATT and the ATE. When the treatment effects vary only randomly across individuals, the expected values of the ATT and ATE will be equal (or homogeneous). Otherwise, when the treatment effect varies at different levels of a baseline characteristic or the propensity score that is a function of one or more such baseline characteristics, it is considered heterogeneous (Harder, Stuart, & Anthony, 2010; Kurth et al., 2006; Wang, Lagakos, Ware, Hunter, & Drazen, 2007) and the ATT and ATE will no longer be expected to be comparable, as will be further discussed in Section 2.4.

In addition to the two estimands, researchers have a choice of method to use to estimate the treatment effect. Previous studies have suggested that approaches other than simply comparing the two matched groups should be used when utilizing matched groups to evaluate the treatment effect. Regression analysis which can also control for remaining covariate differences between matched treated and control groups in observational studies is a typical option (Cochran & Rubin, 1973; Ho et al., 2007; Rubin, 1973; Rubin & Thomas, 2000). This way, the matched data are further adjusted to control for the small remaining pretreatment differences between groups. The combination of conditioning and regression analysis is referred to as the "doubly robust" procedure which produces a consistent effect estimator as long as one of the methods in the combination is correctly specified (Funk, Westreich, Davidian, & Wiesen, 2011; Stuart & Rubin, 2007).

Following the ideas introduced above, both the ATT and ATE can be obtained as the estimated coefficient for the treatment variable in a regression function. The functional form is shown as:

$$Y_i = \gamma_0 + \gamma_1 T_i + \sum_{p=1}^{P} \beta_p X_{pi}, \tag{11}$$

where Y_i is the outcome score for person i (i.e., math achievement), γ_0 is the intercept of the model, $\gamma_1$is the TE estimate, and T_i is the treatment status. The term $\sum_{p=1}^{P} \beta_p X_{pi}$ represents the idea of doubly robust, that is, the linear regression adjustment after conditioning is applied. Although a doubly robust approach is a common method to improve the TE estimates, it was not an interest of this simulation study. Given that the true PS model was known and all pretreatment covariates used for data generation were included in the PS estimation model, doubly robust was unlikely to further explain the variance in the outcome.

Summary of Section 2.2: In this subsection, I introduced what a propensity score is, in what areas it has been applied, the important assumptions that underlie the estimation of causal effects using PS methods, and how to implement these PS methods step-by-step. With this introduction of the basic steps of using the PS methods, the next section, 2.3, explains the theoretical framework of how relative PS distributions could possibly affect the performance of the methods, the central focus of this proposed dissertation study.

2.3 Potential Effect of Propensity Score Distributions

In a non-experimental design where the data are not randomly assigned to the treatment conditions, the treated and control groups may differ systematically with respect to the related pretreatment covariates. As a function of these covariates, the PS estimates may distribute differently for each treatment group. Specifically, assuming the model used to predict probability of treatment ("the PS model") is correct (the functional form is accurate) and informative (the selection into treatment was not random), the PS estimates may differ in a variety of ways including means, variances, skewness, and kurtosis. In terms of the PS density distributions, the propensity scores tend to distribute towards 1 for the treated group and 0 for the control group (see Figure 1), reflecting that the overall probabilities of receiving the treatment are higher for the treated group and lower for individuals in the control group. In short, the differences in the pretreatment covariates between the two groups would lead to different relative PS distributions. Because the relative PS distributions may differ in unlimited forms (due to unlimited patterns of differences in a number of covariates and thus unlimited possibilities of central tendency, variability, skewness, and kurtosis of the associated PS distributions), in this study, the distinction of the relative PS distributions is defined as the *overlap* of the two distributions (i.e.,

the proportion of the intersection of the two distributions divided by the union of the two distributions). A larger distinction between the two PS distributions yields a smaller overlap.

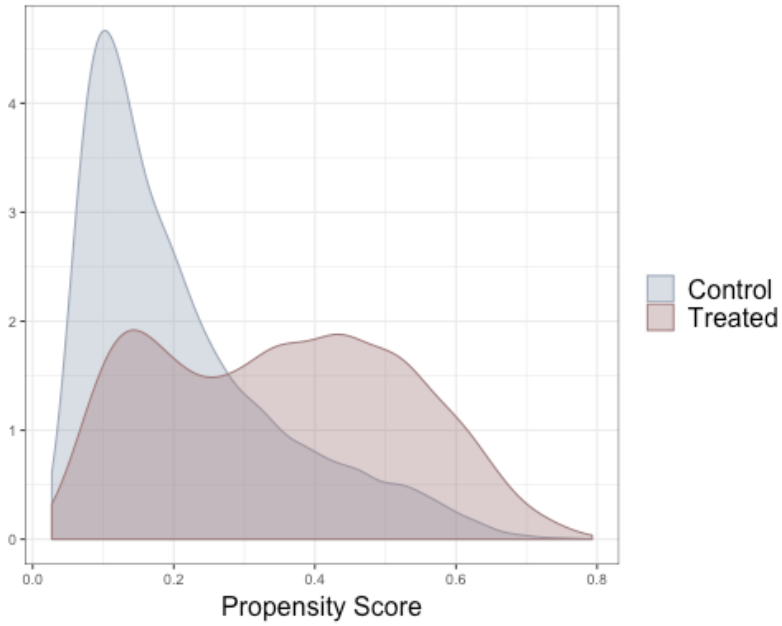


Figure 1. An example of relative PS distributions of the treated and control groups (via probability density functions). In this example, these two distributions differ in multiple ways including mean, variance, skewness, kurtosis, in addition to overlap.

Arguments have been made in the literature that the estimated PS distributions are not meaningful, because the purpose of PS conditioning is to obtain two comparable samples rather than accurate PS estimates (Dugoff et al., 2014; Hahs-Vaughn, 2015). However, good balance is not the ultimate goal of the use of a PS method and does not guarantee that the TE estimates are unbiased. As clearly seen from the Subsection 2.2.3.3, all of the conditioning methods are directly based on the PS estimates. Different PS estimates, or the relative PS distributions (or overlap) between the two groups, accordingly, may affect the performance of each conditioning method.

To illustrate, when there is a relatively large distinction between the two PS distributions, the overlap would be smaller. As a result, for PS matching, more subjects with relatively extreme

propensity scores will be discarded without having appropriate matches and the matched data could be smaller. For PS subclassification, the cut-off propensity scores used for grouping could be more extreme; meanwhile, the PS overlaps between the two groups within each subclass, especially within the highest and the lowest subclasses, could be very small (i.e., few cases for one group and many cases for the other), which could lead to unstable within-subclass estimates, and in turn affect the overall TE estimates for the full sample. When there are few control or treatment cases in a stratum due to extreme propensity scores, strata need to be collapsed and that likely would further add bias to the TE estimates.

With respect to PS weighting, when there are extreme propensity scores, the PS weights could be extremely high. Table 1 shows examples of IPTW and WBO weights (calculated with Equations 6 and 7) for each group at different levels of propensity scores. A concern with PS weighting is that overweighting abnormal subjects could lead to biased and less precise TE estimates (Austin, 2011a). Take the propensity score of 0.99 as an example, the corresponding IPTW weight is 100 for a control subject, indicating that this particular subject represents 100 subjects in the population when estimating the treatment effects. Imagine that this subject happens to be a student who never participated in the math tutoring program but had a superior math grade, his math performance would possibly inflate the average math scores for the control group and lead to underestimation of the treatment effect of that math program. Furthermore, in the case of PS adjustment via weighting, extreme weights that are due to extreme propensity scores may increase the variability of the weights and thereby result in a loss of precision for the TE estimates (Czajka et al., 1992; Kalton & Flores-Cervantes, 2003; Lee, Lessler, & Stuart, 2011). The quantity of the loss of precision is a function of the following variance inflation factor:

$$F = 1 + \left(\frac{\sigma_w}{\mu_w} \right)^2, \tag{12}$$

where σ_w and μ_w are the standard deviation and mean of the weights respectively (Kalton & Flores-Cervantes, 2003; Kish, 1992). As clearly seen from Equation 12, more extreme weights tend to inflate the ratio σ_w/μ_w (also referred to as the coefficient of variation), and in turn, lead to a loss in precision. In summary, for PS weighting, the possible changes in both accuracy and precision due to potential different extreme weights based on the relative PS distributions may contribute to treatment effects of different qualities.

Table 1

IPTW and WBO Weights at Different Levels of Propensity Scores

Propensity Scores	IPTW Weight		WBO Weight	
	Treated	Control	Treated	Control
0.99	1.01	100.00	1.00	99.00
0.90	1.11	10.00	1.00	9.00
0.80	1.25	5.00	1.00	4.00
0.70	1.43	3.33	1.00	2.33
0.60	1.67	2.50	1.00	1.50
0.50	2.00	2.00	1.00	1.00
0.40	2.50	1.67	1.00	0.67
0.30	3.33	1.43	1.00	0.43
0.20	5.00	1.25	1.00	0.25
0.10	10.00	1.11	1.00	0.11
0.01	100.00	1.01	1.00	0.01

Researchers have developed two popular techniques, weight stabilization and trimming, to accommodate the undesirable effects of extreme weights. These two techniques are different in purpose and implementation. The way to stabilize weights is to multiply a constant, that is, the mean of the probabilities of being assigned to the corresponding group, to the original PS weights (Robins, 1998, 1999; Robins, Hernán, & Brumback, 2000). The equations for weight stabilization for IPTW and WBO are respectively shown in Equations 13 and 14.

$$sw_{IPTW.i} = \frac{\sum_{i=1}^{N_T}\left[T_i e_i + (1-T_i)(1-e_i)\right]}{N_T} \times w_{\text{IPTW}.i}, \quad (13)$$

$$sw_{WBO.i} = \frac{\sum_{i=1}^{N_T}\left[T_i e_i + (1-T_i)(1-e_i)\right]}{N_T} \times w_{\text{WBO}.i}, \quad (14)$$

where $sw_{IPTW.i}$ and $sw_{WBO.i}$ represent the stabilized IPTW and WBO weights, $w_{IPTW.i}$ and $w_{WBO.i}$; N_T is the size of group T, which is either the treated or the control group; T_i is the treatment status for person i; and e_i is the PS estimate for person i. The calculations for $w_{IPTW.i}$ and $w_{WBO.i}$ are shown in Equations 6 and 7.

Because stabilization does not change the weighted means for each group, it does not affect the point estimate. The purpose of implementing stabilization, then, is to reduce the variance (or to improve the precision) of the estimates. Robins and colleagues (1998, 1999, & 2000) also found that when the PS model is accurately specified, stabilization will not change the variance either and thus is not useful.

Weight trimming is a technique often used in survey analysis to truncate the extreme weights to improve the precision of the outcome estimates. Although a decrease in sampling error can be expected with the use of trimming, this technique increases the potential for bias in the estimates, because the weighted cases after trimming are no longer perfectly representative of the original population (Potter, 1990). The weight trimming technique has been recently introduced into PS studies (Lee, et al., 2011). Lee et al. conducted a simulation study to investigate if the benefits of weight trimming found in the survey sampling setting also apply to the PS setting and whether the benefits vary depending on PS estimation methods (i.e., logistic regression vs. non parametric methods such as boosted classification trees and random forests).

The authors found that with weight trimming following logistic regression, the accuracy and precision of final parameter estimates were improved; however, the benefit did not apply to any nonparametric PS estimation. They also indicated that although weight trimming can improve causal inferences in some settings, the research should always focus on improving the generated weights (e.g., via specifying a more accurate PS model) rather than relying on this ad-hoc method of trimming. The weight trimming technique was still applied in this study to provide comprehensive information for those who are interested in this technique.

A popular way of performing trimming is to use percentile cutpoints (Cole & Hernán, 2008; Lee et al., 2011), that is, setting a percentile cutpoint as the maximum value of the weights. Taking the 90^{th} percentile as an example, any weights that are greater than the 90^{th} percentile are set equal to the 90^{th} percentile. This way, the extremely high are trimmed downwards and will not inflate to an unexpected degree. As the optimal level of trimming is difficult to determine, one can set a range of cutpoints ranging from the 50^{th} to the 99^{th} percentiles and report all results. In addition to setting percentile cutpoints, one can also set the original PS weights to a range based on the raw weight values (i.e., no more than 10; Harder et al., 2010).

According to the above description, the undesirable effect of relative distributions regarding extreme weights on the precision of the TE estimates can be addressed, in part, by the application of weight stabilization and trimming. However, the potential bias caused by different relative PS distributions, as well as the further biased TE estimates from weight trimming have been rarely discussed in the PS literature.

With the discussion of the potential effects of PS relative distributions, the primary focus of this study was to explore how the selection of a specific PS conditioning method might be affected by different levels of PS relative distributions. This issue was investigated via a

simulation study in which different relative PS distributions (in terms of different sizes of overlap, in combination with mean, variance, skewness, and kurtosis) were manipulated. By comparing the accuracy and precision of the TE estimates obtained using different PS conditioning methods, I was able find out which PS conditioning method had the best performance under which relative PS distribution condition. As a brief summary, PS matching and subclassification are likely to be sensitive to the relative PS distributions in certain conditions, while PS weighting without trimming tends to be more robust across a variety of data conditions. The results are discussed in more detail in Chapter 4.

2.4 Potential Effect of Heterogeneous Treatment Effects

Section 2.2.3 reviewed that ATT and ATE are both average treatment effects across individuals. The treatment effect is considered homogeneous when the effects vary only randomly for all individuals. In practice, however, there is a possibility that the treatment effects vary systematically depending on certain covariates. For example, subjects from different pretreatment subgroups may have different treatment effects. Continuing the example from the last section, the effect of a math tutoring program, the program may have a larger effect on students who spend a longer time on studying, which is also a covariate in the PS model. In this case, when the treatment assignment interacts with any covariate of the PS model (or the PS estimates that are based on the related covariates), the treatment effects will vary across subjects, in other words, are heterogeneous. That said, the ATT and ATE are very likely to differ when heterogeneous treatment effects exist, because they represent different populations, in which the distributions of the treatment effect-related covariates may also differ.

Section 2.3 explains how relative PS distributions can affect the conditioned data and thus the TE estimates. As was introduced in Section 2.2.3.5, when the TE estimates are

heterogeneous, individuals with different PS estimates will have different individual treatment effects (ITE) given that the ITEs are associated with one or more pretreatment covariates. The heterogeneous treatment effects could further affect the TE estimates via interacting with the relative PS distributions. For PS matching, relative PS distributions affect TE estimates via changing which and how many treated cases are discarded; heterogeneous effects could further affect the TE estimate due to the fact that the cases discarded may have different ITEs. When it comes to PS weighting, the situation is more complicated. Consider the ATT estimation using WBO weights as an example, imagine that the treated and control groups each have 11 subjects with the 11 propensity scores shown in Table 1, while the WBO weights for those in the treated group are all 1, the relative weights (i.e., weight/ sum weight) for the control subjects vary substantially (shown in Figure 2). When the ITEs are independent of the PS estimates, the ATT estimate will just be the weighted mean difference of the two groups. If the ITEs are heterogeneous, in other words, correlated with one or more pretreatment covariates, the ITEs would increase monotonically along with the PS estimates (also shown in Figure 2). The overall ATT estimate, then, is the function of the association between the relative weights and the heterogeneous ITEs. In a condition with large relative distributions of propensity scores, for example, more subjects in the control group will have large weights and thus larger OVERALL relative standing in the treated population; if these large weights are further associated with greater ITEs, the overall ATT estimate will be inflated compared to the condition of homogeneous treatment effects. Similar effects of heterogeneous treatment effects will apply to IPTW weighting, as well as subclassification, which has the combined features of PS matching and PS weighting. Therefore, different levels of heterogeneity in treatment effects will also be manipulated in the study, to facilitate the exploration of the relative PS distributions.

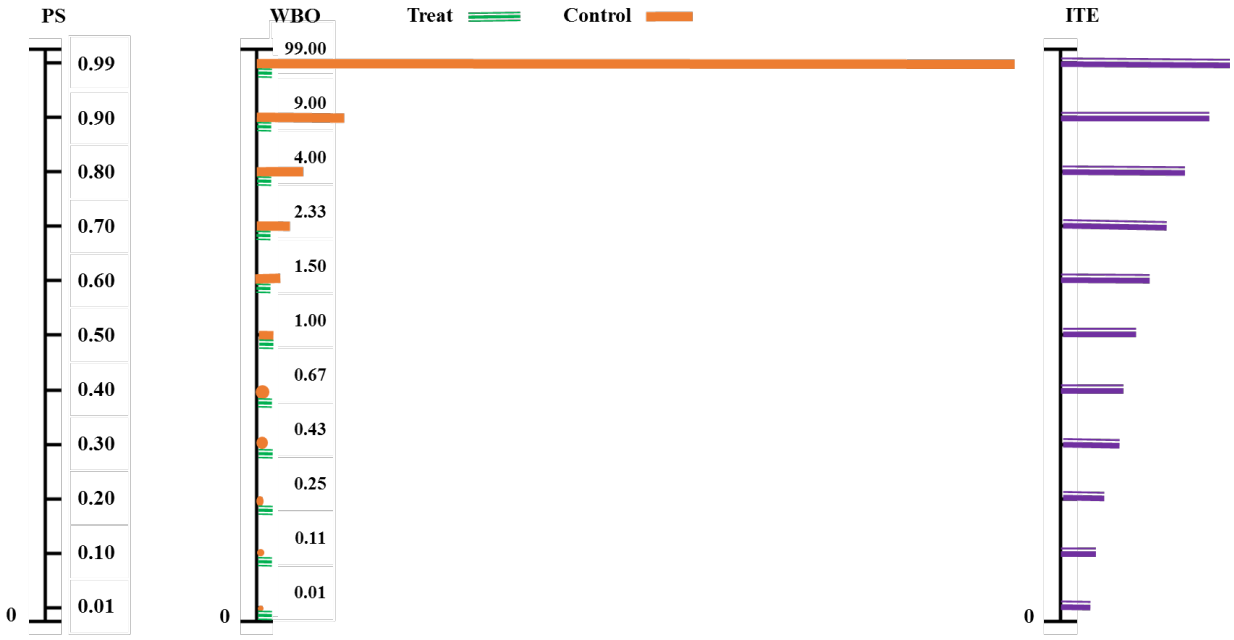


Figure 2. Illustration of the possible effect of heterogeneous treatment effects.

2.5 Research Questions

The purpose of this dissertation was to investigate whether, and in what way, the relative PS distributions between the treated and control groups affect the performance of each PS conditioning method. The effects of interest were examined in terms of the accuracy and precision of the TE estimates, by answering the following methodological questions via a simulation:

(1) Do the relative PS distributions between the treated and control groups affect the quality of the TE estimates obtained via different PS conditioning methods?

(2) Does the relation between the relative distributions and performance differ across levels of TE heterogeneity?

The answers to these questions would have the potential to affect behaviors of researchers who use PS methods – they might first be encouraged to evaluate the differential PS distribution

across groups, and then choose a method that works best for that kind of differential. Findings from this study thus will add to the literature on the implementation of the PS method – an extra step of checking the relative PS distribution and select the best PS conditioning method, in addition to the traditional five steps, might be necessary to make more accurate and precise causal inference.

While the simulation study was to answer the above research questions, as a demonstration of the issues involved in addressing the importance of PS distributions across groups, an empirical data analysis was conducted to investigate the relation between having home access to computers and first grade students' math achievement using data selected from the Early Childhood Longitudinal Study Kindergarten Class of 2010-11 (ECLS-K:2011; Tourangeau et al., 2015). It did not only serve as an illustration of the proposed six-step PS procedure, obtaining an estimate of the treatment effect of having a home computer may also be desirable for education policy makers. In Chapter 3, I present the simulation design to evaluate the two research questions and, in Chapter 4, I interpret the results. Chapters 5 and 6 respectively discuss the design and results of the empirical analysis.

Chapter 3. Simulation Design

The primary goal of this study was to investigate whether the relative PS distributions between the treated and control groups affect the quality of the TE estimates obtained via different PS conditioning methods. Another question the study was trying to answer was whether the relation between the relative distributions and performance of the PS conditioning methods would be different across levels of TE heterogeneity. The answers to these questions were evaluated through a simulation study.

In this chapter, I begin by introducing briefly how the major manipulated factors were selected, followed by describing how each of the study factors was generated in more detail. I then specify the implementation of two major steps involved in a typical PS method, including PS estimation and the PS conditioning methods used in this study, as well as the selection of the TE model. Next, I provide a summary of the simulation conditions, followed by measures of performance that were compared to answer the research questions.

3.1 Important Simulation Factors

The simulation mainly manipulated three fully-crossed factors: differences in the relative PS distributions, the level of heterogeneity in treatment effects, and sample size. First, the effect of PS distributions on the conditioning performance was the focus of the study and thus the level of relative PS distributions, in terms of varying overlap, was manipulated. The four empirically-defined relative distributions that were manipulated in the study are shown in Figure 3. The overlap quantities (.61, 0.35, 0.20, and 0.32 from left to right) were calculated as the proportion of the intersection of the two PS distributions over the union of the two distributions, based on super populations with a sample size of 50,000 for both groups, with homogeneous treatment effects. The calculation was done via integration, using the R package of "*sfsmisc*". The overlap

shows how similar the PS distributions are between the two groups, which further indicates how comparable the two groups are before PS adjustment. With smaller overlap (or greater the difference in the PS distributions), more pretreatment adjustment is needed to get the data ready for TE esimation. The hypothesis here was that some PS conditioning methods may perform better than others when more pretreatment adjustment is needed.

The second factor manipulated was the level of heterogeneity in treatment effects. This factor was related to the two popular causal estimands, ATT and ATE. As was indicated in Section 2.2.3.5, the treatment effects are considered homogeneous when they are independent of the propensity scores, that is, the expected values of ATT and ATE are equal. When the treatment effects are heterogeneous, they must be correlated with one or more of the pretreatment covariates and would possibly interact with the relative PS distributions, which may in turn, affect the performance of the PS methods. Therefore, both homogeneous and heterogeneous treatment effects were simulated to further understand the relations between relative PS distributions on the performance of PS conditioning methods. Three levels of heterogeneity were manipulated via an interaction of the treatment exposure variable and a pretreatment covariate in the TE model. The heterogeneity level of treatment effects was 0 (homogeneity) when this correlation was set 0.

Sample size was the third factor manipulated in the simulation study. This sample size factor is mostly likely to affect PS matching and subclassification. For PS matching, depending on the specific matching technique (e.g., 1:1 matching with a caliper), not all treated cases can find a match from the control group. Therefore, the sample size, particularly the relative sample sizes in the two groups, may affect the matching rate and thus may furhter affect the accuracy of the treatment effect. For subclassification, the relative numbers of cases falling in each subclass

may change how the treatment effects are weighted across subclasses and further affect the accuracy and precision.

3.2 Data Generation

Considering the conditions discussed above, a simulation was designed with population data generated following the three fully-crossed factors:

(1) PS distributions with small, medium, and large differences between the two groups (with the overlap of 0.61, 0.35, and 0.20 calculated as the proportion of the intersection of the two relative PS distributions over the union of the two distributions). In all these three conditions, the PS esimates in both groups have a full coverage from above 0 to below 1. Another "truncated" form of relative PS distributions was generated to reflect the situation when the PS estimates in the two groups have different coverage (with a simulated overlap of 0.32). These four PS distribution conditions are visually shown in Figure 3; the statistical features for each distribution are presented in Table 2.

(2) Heterogeneity of treatment effects at zero (i.e., homogeneity), slight, and substantial levels. The heterogeneous treatment effects were generated by increasing the interaction effect between the treatment status and one of the pretreatment effects in the outcome generation model. Details on how these effects were quantified are illustrated in the following along with the introduction to the true PS and TE models. When heterogeneity exists (i.e., slight or substantial), the ATT and ATE are different when the heterogeneity is not zero.

Table 2

Descriptives of the Simulated PS Relative Distributions

Relative Distribution	Overlap	Mean		Variance		Skewness		Kurtosis	
		Treat	Control	Treat	Control	Treat	Control	Treat	Control
Small	0.61	0.56	0.37	0.09	0.08	-0.24	0.50	1.75	2.00
Medium	0.35	0.57	0.21	0.09	0.05	-0.26	1.37	1.76	4.06
Large	0.20	0.57	0.10	0.09	0.02	-0.25	2.53	1.76	9.96
Truncated	0.32	0.86	0.60	0.02	0.05	-1.81	-0.36	6.53	2.13

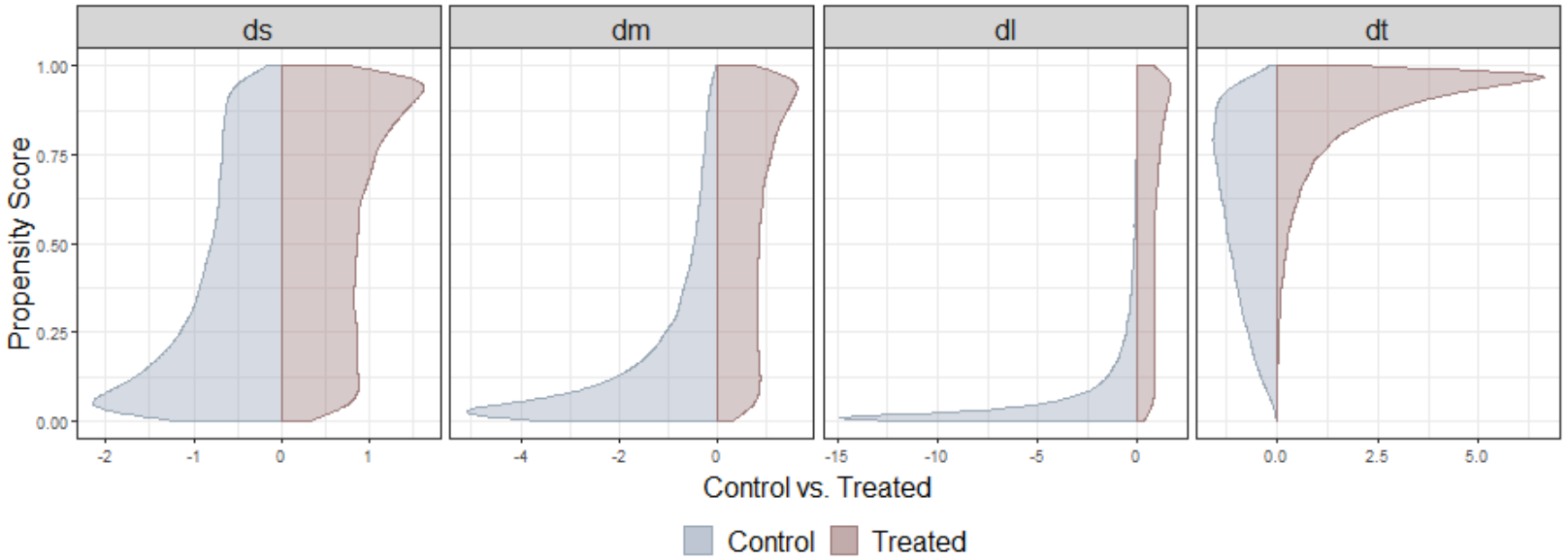


Figure 3. The relative PS distributions of the treated and control groups; ds – small PS relative distribution, dm – medium PS relative distribution, dl – large PS relative distribution; dt – truncated PS relative distribution.

(3) Small, medium, and large sample sizes for the control group. Specifically, 500, 1500, and 5000 cases for the control group and correspondingly 500, 500, and 500 cases for the treatment group were generated.

Because the true PS model is usually unknown with real data, PS model misspecification is a common issue in practice. In addition to the three primary factors introduced above, the simulation also generated two types of PS population data as another factor. For the two population settings, the same PS estimation model would be considered correct and incorrect specifications respectively. The population setting where the PS model was correctly specified is represented as Scenario A and that with PS misspecification is Scanrio B. Scenario A data generation followed, in part, the additivity and linearity model introduced in Setoguchi, et al. (2008) which was further applied in Lee et al (2010). In addition to the Scenario A framework with additivity and linearity (i.e., main effects only), a second set of data with moderate non-additivity and non-linearity (i.e., ten two-way interaction terms and three quadratic terms) were generated in Scenario B (Lee et al., 2010; Setoguchi et al., 2008). The data generation models are specified in Equations 15 and 16.

$$\text{logit}(e \mid T = 1) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon, \quad (15)$$

where $\beta_1 - \beta_7$ are the simulated coefficients for the true PS model in Scenario A, where $\beta_1 = \beta_5 = 0.80$, $\beta_2 = 0.25$, $\beta_3 = 0.60$, $\beta_4 = 0.40$, $\beta_6 = 0.50$, and $\beta_7 = 0.70$; ε is the random error term; and $X_1 - X_7$ are vectors of random normal variables with mean of 0 (for the small relative PS distributions condition) and standard deviation of 1. As the level of relative PS distributions moved up by one category, the means of each covariate in the treated group were decreased by an effect size (measured by Cohen's *d*) of 0.3. The conceptual model for this

generation is shown in Figure 4, with correlations illustrated in the figure. Some of the continuous variables then were converted to binary variables and thus the actual correlations of the generated variable values were smaller. The correlations among the variables were generated to reflect the complexity in research data in the real world. For example, in an educational setting, X2 represents for mother's education level for a student and X6 is whether the student goes to a private school. These two variables are likely to be positively correlated naturally in the real world for the reason that the education level may directly affect the household income, living area, expectation on the children, etc.

The data for Scenario B, with moderate non-additivity and non-linearity, were generated following Equation 16:

$$\begin{aligned}&\text{logit}(e \mid T=1) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \\ &\beta_2 X_2^{\,2} + \beta_4 X_4^{\,2} + \beta_7 X_7^{\,2} + \\ &(0.5)\beta_1 X_1 X_3 + (0.7)\beta_2 X_2 X_4 + (0.5)\beta_3 X_3 X_5 + (0.7)\beta_4 X_4 X_6 + (0.5)\beta_5 X_5 X_7 + \\ &(0.5)\beta_1 X_1 X_6 + (0.7)\beta_2 X_2 X_3 + (0.5)\beta_4 X_4 X_5 + (0.5)\beta_5 X_5 X_6 + \varepsilon\end{aligned}, \qquad (16)$$

where the variables and coefficient remain the same as in Equation 16. The parameters used for generating the covariates were consistent with those in Scenario A (see Equation 15).

When fitting the non-additive and non-linear data generated in Scenario B with the same linear logistic regression PS model as used for the additive and linear data in Scenario A, ignoring the interaction and quadratic terms, the fitted PS model would be considered misspecified. This way, both correct and misspecified PS models were examined in combination with the proposed PS conditioning methods. The results revealed how PS misspecification could possibly interact with the PS relative distributions and homogeneous treatment effects, and in turn, affect the performance of each PS conditioning method.

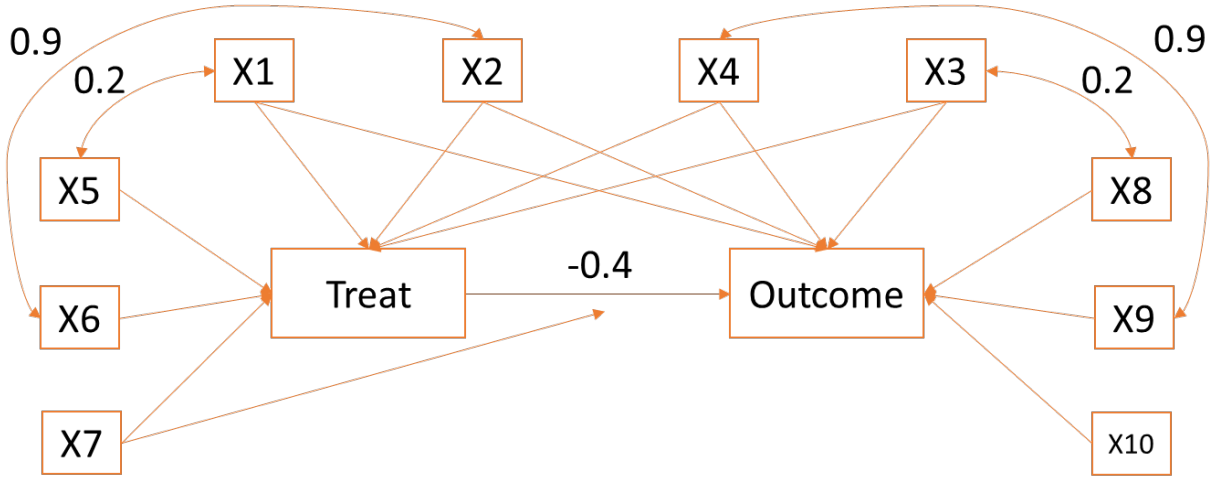


Figure 4. The data generation model for the simulation study (Scenario A). The framework was originally created in Setoguchi et al. (2008), and was further applied in Lee et al. (2010) and Austin (2012). X1, X3, X5, X6, X8 and X9 are dichotomized variables, while the others are random normal variables. $Corr\left(X_1, X_5\right) = Corr\left(X_3, X_8\right) = 0.2$,

$Corr\left(X_2, X_6\right) = Corr\left(X_4, X_9\right) = 0.9$.

In addition to the PS generation model, the model used to generate outcome scores is expressed as:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10} + \gamma_0 T + \gamma_1 X_7 T + \varepsilon, \qquad (17)$$

where $X_1 - X_4$ and $X_8 - X_{10}$ are the same covariates from the PS generation model; $\alpha_0 - \alpha_7$ are the simulated coefficients for the for the true TE model, where $\alpha_0 = -3.85$, $\alpha_1 = 0.30$, $\alpha_2 = -0.36$, $\alpha_3 = -0.73$, $\alpha_4 = -0.20$, $\alpha_5 = -0.71$, $\alpha_6 = -0.19$, and $\alpha_7 = 0.26$; T is the treatment exposure and thus the simulated baseline treatment effect was $\gamma_0 = -0.40$ when there was no heterogeneity. The homogeneous treatment effect was generated by adding an interaction effect of γ_1 between the treatment status T and one of the covariates X_7. To simulate three different sizes of the heterogeneous treatment effects, γ_1 was set to be 0 to reflect zero

heterogeneity (homogeneous treatment effect), -0.40 and -0.80 for the slight and substantial heterogeneity. Therefore, the real treatment effects (homogeneous or heterogeneous) can be expressed as:

$$\gamma_{heter} = \gamma_0 + \gamma_1 X_7 . \tag{18}$$

The actual sizes of the heterogeneous treatment effects were measured via coefficient of variation cv_{heter}, a standardized measure of dispersion of the heterogeneous treatment effects distribution defined as the ratio of the standard deviation δ_{heter} to the mean μ_{heter}:

$$cv_{heter} = \frac{\delta_{heter}}{\mu_{heter}} . \tag{19}$$

The values of cv_{heter} were respectively 0, -1.68, and -7.54 for zero (i.e., homogeneous), slight, and substantial heterogeneous treatment effects based on super populations of 50,000 cases in both treated and control groups, under the condition of large relative PS distributions.

3.3 PS Estimation

Following data generation, propensity scores were estimated via a logistic regression before undertaking any PS conditioning techniques. The data in both scenarios were fit via the PS estimation model expressed in Equation 20:

$$logit\left(\hat{e} \mid T = 1\right) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6 + \hat{\beta}_7 X_7, \tag{20}$$

where $\hat{\beta}_0$ - $\hat{\beta}_7$ are the estimated coefficients for each of the pretreatment covariates $X_1 - X_7$. As was indicated in Section 3.2, the PS model was cosidered mispecified in Scenario B where moderate non-additivity and non-linearity were generated.

3.4 PS Conditioning Methods and Balance Check

Once the PS estimates were available, the selection bias between the two treatment groups were adjusted by different PS conditioning methods. In this study, I compared the TE estimates after conditioning with PS matching, subclassification, and PS weighting (i.e., IPTW and WBO). Each of the choices made for these conditioning methods is described as follows.

3.4.1 Matching

Among the various matching options, 1:1 nearest neighbor matching without replacement with a caliper of 0.25 (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1984) was used as an example due to its popularity and simplicity in practice. As unmatched cases were removed in the PS matching process, this technique was used to obtain an estimate of the ATT.

3.4.2 Subclassification

The subclassification technique was based on 5 subclasses identified by the quintiles of the overall propensity scores to remove at least 90% of the bias (Rosenbaum & Rubin, 1984). Both estimates of the ATT and ATE were calculated with subclassification following the corresponding methods of weighting TE estimates for each subclass introduced in Section 2.2.3.3. If there were not sufficient numbers of treated or control cases in any subclass, for example, 0 control in the first stratum, I collapsed the adjacent strata for the simulation and calculated the TE utilizing fewer than 5 subclasses.

3.4.3 PS weighting

Similarly, for PS weighting, IPTW and WBO were applied to obtain estimates of both the ATE and ATT. Weight stabilization was not used in combination with PS weighting in this simulation study, because in the scenario of the primary interest of the study (Scenario A), PS specification in Scenario A the PS model were accurately specified and thereby stabilization

would not be useful in improving the precision of the estimates (Robins and colleagues, 1998, 1999, and 2000). As for trimming, I set a lower bound of 0.10 and an upper bound of 10 as the cutpoints for the raw WBO and IPTW weights following Harder et al. (2010). In particular, any weight that was greater than 10 or smaller than 0.10 was set to 10 or 0.10 instead, respectively. Other options, such as setting percentile cutpoints as the maximum value of the weights (Cole & Hernán, 2008; Lee et al., 2011), were not considered in this study. This is because trimming was only implemented to demonstrate an option for PS application, and thus finding and using the optimal level of trimming was not the interest of the study.

3.4.4 Balance check

Following the application of these PS conditioning methods, balance in terms of SMD (Equation 8) was computed and compared to the criterion of 0.25. In an applied setting, PS practitioners usually keep tweaking the PS models until the point at which the balance measures meet the criterion. For this simulation study, the balance metric was only used as a *supplemental* index for the examination of the relations between the performance of the PS conditioning methods and relative PS distributions. That said, I moved to TE estimation for all conditions, regardless of the quality of balance.

3.5 TE Estimation Model

With the comparable samples were obtained via each of the above PS conditioning techniques, the treatment effect, then, was estimated as the function of the treatment. Doubly robust was not considered for this simulation study because all related covariates were known and included in PS estimation, and thus including them in the TE estimation model would unlikely in further explaining the remaining variance in the outcome.

Additionally, the simulated interaction term between the treatment (T) and the covariate, X_7 , was not considered either, because the interaction might not be anticipated by (nor of interest to) the researcher. Therefore, the fitted TE model can be expressed as:

$$\hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 w_i T_i, \tag{21}$$

where $\hat{\gamma}_0$ is the intercept and $\hat{\gamma}_1$ represents the TE estimate; w_i is the weight for each subject. The weight (w_i) was 1 for all subjects for matching and subclassification; it represents the corresponding WBO or IPTW weights for each subject in the PS weighting methods.
For subclassification, the final TE estimate, as well as the SE estimate, were calculated as the weighted mean across all subclasses. Details on how weighting was done across subclasses for ATT and ATE are described in detail in Section 2.2.3.3. For PS weighting, weighted means were compared instead between the two groups. Because the observations were weighted in the PS weighting methods, the standard errors were computed using the sandwich variance estimator, which was robust to non-constant residual variance.

3.6 Summary of Simulation Conditions

To summarize the controlled conditions, the designed study had 72 cells generated by the 3 primary study factors(i.e., relative PS distributions, heterogeneity in treatment effect, and sample size) and 1 secondary factor (PS specification). Within each cell, 8 PS conditioning methods (including a naïve model without adjustment for selection bias) were implemented to estimate ATE or ATT. All these simulation conditions are summarized in Table 3. The data generation, PS estimation and conditioning, and TE estimation were repeated using 250 Monte Carlo replications within each cell. . All data generation and analyses in this dissertation were conducted using R. The *MatchIt* and *twang* packages were used for the PS for the implementation of the PS methods.

Table 3

Summary of Simulation Conditions

		large rel. PS dist.			medium rel. PS dist.			small rel. PS dist.			truncated rel. PS dist.		
		heterogeneity			heterogeneity			heterogeneity			heterogeneity		
		substantial	slight	zero	substantial	slight	zero	substantial	slight	zero	substantial	slight	zero
sample size	small	Method 0: naïve method using unconditioned sample without any selection adjustment. Method 1: 1:1 nearest neighbor matching without replacement with a caliper of 0.25. Method 2: subclassification with five PS subclasses to estimate ATT. Method 3: weighting by the original WBO weights to estimate ATT. Method 4: Weighting by the trimmed WBO weights to estimate ATT. Method 5: subclassification with five PS subclasses to estimate ATE. Method 6: weighting by the original IPTW weights to estimate ATE. Method 7: weighting by the trimmed IPTW weights to estimate ATE.											
	medium												
	large												

Note. This table summizes both scenarios (A and B).

3.7 Measures of Performance

The quality of the TE estimation (for both ATE and ATT) under each condition was evaluated in terms of accuracy and precision of the TE estimates via relative bias, empirical (i.e., true) standard error, and relative bias in the SE estimate. The definitions of these three measures are presented in Equations 22, 23, and 24.

$$\theta_{REL.BIAS} = \frac{1}{B}\sum_{r=1}^{B}(\hat{\theta}_r - \theta_{TRUE}) / \theta_{TRUE}\,, \tag{22}$$

where $\theta_{REL.BIAS}$ is the relative bias of the estimated treatment effects across all B replications, $\hat{\theta}_r$ is the TE estimate for the r^{th} replication, and θ_{TRUE} is the simulated true treatment effect, which is the function of the simulated baseline effect, interaction coefficient between treatment and a covariate, and that covariate itself (Equation 18).

$$SE(\hat{\theta}) = \sqrt{(\frac{1}{B-1})\sum_{r=1}^{B}(\hat{\theta}_r - \bar{\hat{\theta}})^2}\,, \tag{23}$$

where $SE\left(\hat{\theta}\right)$ is the empirical (true) standard error of the estimated treatment effects $\hat{\theta}$, $\bar{\hat{\theta}}$ is the average of the TE estimates across all replications, and the others are as above.

$$SE(\hat{\theta})_{REL.BIAS} = \frac{1}{B}\sum_{r=1}^{B}(SE(\hat{\theta}_r) - SE(\hat{\theta})) / SE(\hat{\theta})\,, \tag{24}$$

where $SE\left(\hat{\theta}\right)_{REL.BIAS}$ is the relative bias in the SE estimate, $SE\left(\hat{\theta}_r\right)$ is the SE of the treatment effect for the r^{th} replication; the others are as above.

In the results, the relative bias indicates how accurate the TE estimates are across all conditions. The empirical standard error shows how precise the TE estimates are via each PS conditioning method, while the relative bias in SE estimate measures whether the SE estimates

are trustworthy. The two SE metrics would respectively demonstrate how large the actual sampling error was and whether we could trust the estimated SE that appeared in the output as being a good estimate of that SE. Among the three performance metrics, the relative bias in TE and SE estimates were the primary metrics used to decide whether the results were valuable, while the empirical SE estimates could tell how precise the results were when the SE estimates were accurate. Therefore, performance thresholds were only defined for the two relative bias metrics.

Hoogland and Boomsma (1998) suggested that the acceptable relative bias of parameter estimates is less than 5% and the acceptable relative bias of SE estimates is smaller than 10%. In addition to these acceptable thresholds (interpreted as "*good to use*" in the rest of the paper), another category of "*use with caution*" was added between "*good to use*" and "*untrustworthy*" to define the results that can be "*used with caution.*" This transitioning category allows practitioners to further understand how desirable/undesirable their results are. The definitions of each category for each of the two metrics are defined in Table 4. According to the thresholds, the results from each PS conditioning method for all data conditions are presented in Chapter 4 as guideline tables to help practitioners make decisions. Although the Hoodland and Boomsma standard is not ideal for all settings, it is the most popular criterion in literature so far. Other options can be found in Bradley (1978) and Muthén, Kapaln, and Hollis (1987).

Table 4

Guidelines for Results Thresholds

	Relative bias in TE estimate	Relative bias in SE estimate
Good to use	< 5%	< 10%
Use with caution	>= 5% &< 10%	>= 10% &< 15%
Untrustworthy	>= 10%	>= 15%

This chapter has elaborated on the design of the simulation study, including the importance and levels of the four study factors for data generation, that is, relative PS distributions, possible heterogeneous treatment effects, sample size, and PS population models to reflect PS specification. Following the data generation section, the approaches for the analyses of the simulated data were adapted from the five-step procedure introduced in Section 2.2.3. The outcome measures of the simulated study were then discussed. The three measures of performance were compared across the between-cell factors to find out (1) how the performance of PS conditioning methods differ in getting accurate and precise TE estimates across different relative PS distributions, and (2) how this pattern was interacted by different levels of TE heterogeneity, sample size, and PS model specification. The results from this simulation study (discussed in Chapter 4) not only provided guidance for researchers and practitioners about which PS conditioning methods to choose under which data settings, but also demonstrated how to implement PS distribution checking and PS conditioning method selection as an extra step in an empirical setting as further described in Chapter 5.

Chapter 4. Simulation Results

This chapter summarizes and analyzes the results of the simulation study. The following sections will interpret how each of the PS conditioning methods performed under the simulated data conditions. In particular, this chapter summarizes the balance, relative bias in TE and SE estimates, and the precision of each PS conditioning method. As Scenarios A and B (i.e., PS estimation with correct and incorrect specification) demonstrated similar patterns in the results, the following interpretation only focuses on Scenario A. The results for Scenario B are attached in the Appendices as supplemental information. Given the interactions between the manipulated simulation factors, interpreting the effects directly was challenging. Instead, the individual treatment effects (ITEs) are introduced to unpack the effects of PS relative distributions, treatment effect heterogeneity, and sample size on the performance of some of the PS conditioning methods. The averaged ITEs at different propensity score categories for three example conditions (i.e., small sample size in the control group and large difference in PS relative distributions with homogeneous, medium and large homogeneous treatment effects) are presented in Table 7 as reference.

4.1 PS matching

PS matching is a typical method used to estimate the ATT. This section interprets the results for PS matching with respect to balance of the covariates after matching, as well as the accuracy and the precision of ATT estimates based on the conditioned samples produced via 1:1 nearest neighbor matching with caliper (details explained in Chapter 3).

4.1.1 Balance

The balance results for PS matching and other PS conditioning methods for Scenario A is shown in Figure 5. The balance values were robust to heterogeneity in treatment effects because

the heterogeneous treatment effects were generated via adding an interaction effect between the treatment status and one of the covariates to the true TE model. In other words, heterogeneity did not affect the characteristics (e.g., distributions) of the covariates across treatment assignment groups. Therefore, the balance estimates presented in Figure 5 are only for the conditions with a homogeneous treatment effect.

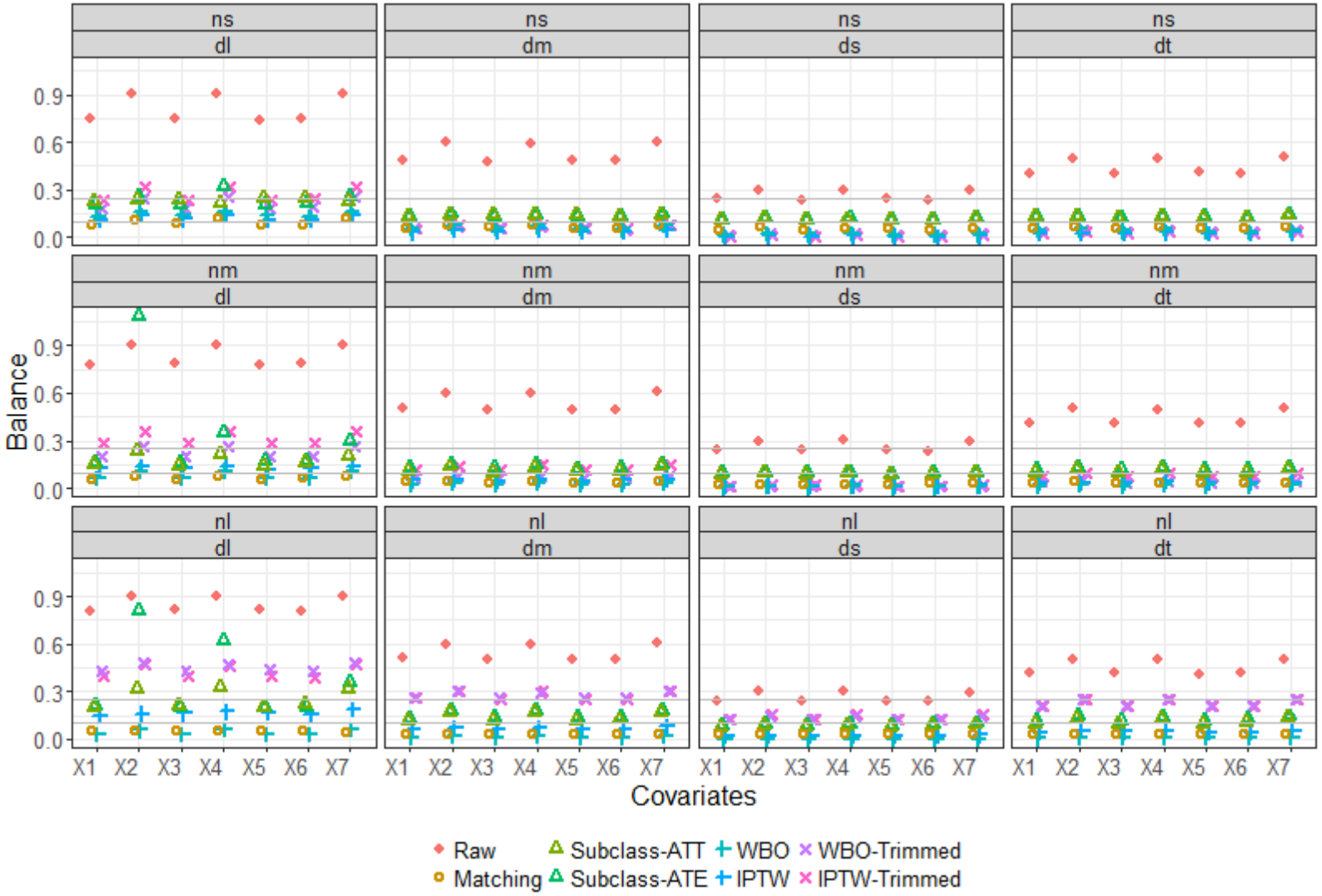


Figure 5. Standardized mean difference (SMD) for all PS conditioning models (Scenario A). The balance plot for scenario B follows the same pattern and is presented in Appendix A. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group.

In general, PS matching produced very good balance between the treatment and control groups across all covariates and conditions manipulated in the simulation study. The SMD values were almost always below the strict criterion of 0.1, except for some covariates in the condition characterized by large PS relative distribution and small control group sample size. The results were similar in Scenario B when the PS score model was misspecified (plot presented in Appendix A).

4.1.2 Accuracy of ATT estimation

PS matching was sensitive to the relative difference in PS distributions across treated and control groups, that is, higher relative bias was associated with larger PS relative distributions, especially with more heterogeneity in treatment effects. However, as the sample size in the control group increased, in general, the bias in the TE estimates decreased (Figure 6, Table 5). Although it would be ideal for PS matching to find matches for all subjects in the treatment group, cases may be discarded from the treatment group depending on whether there is sufficient overlap between the treated and control groups. The smaller the overlap, the more subjects in the treatment group will be discarded. Losing any treated cases would result in some degree of bias in the TE estimates, as the remaining treated cases are no longer fully representative of the overall treated population. In the current simulated conditions, the overlap between the two groups tended to be smaller with larger PS relative distributions, and thus more cases were discarded which in turn, resulted in more bias in the TE estimates. The average matching rates for each simulated condition are presented in Table 6.

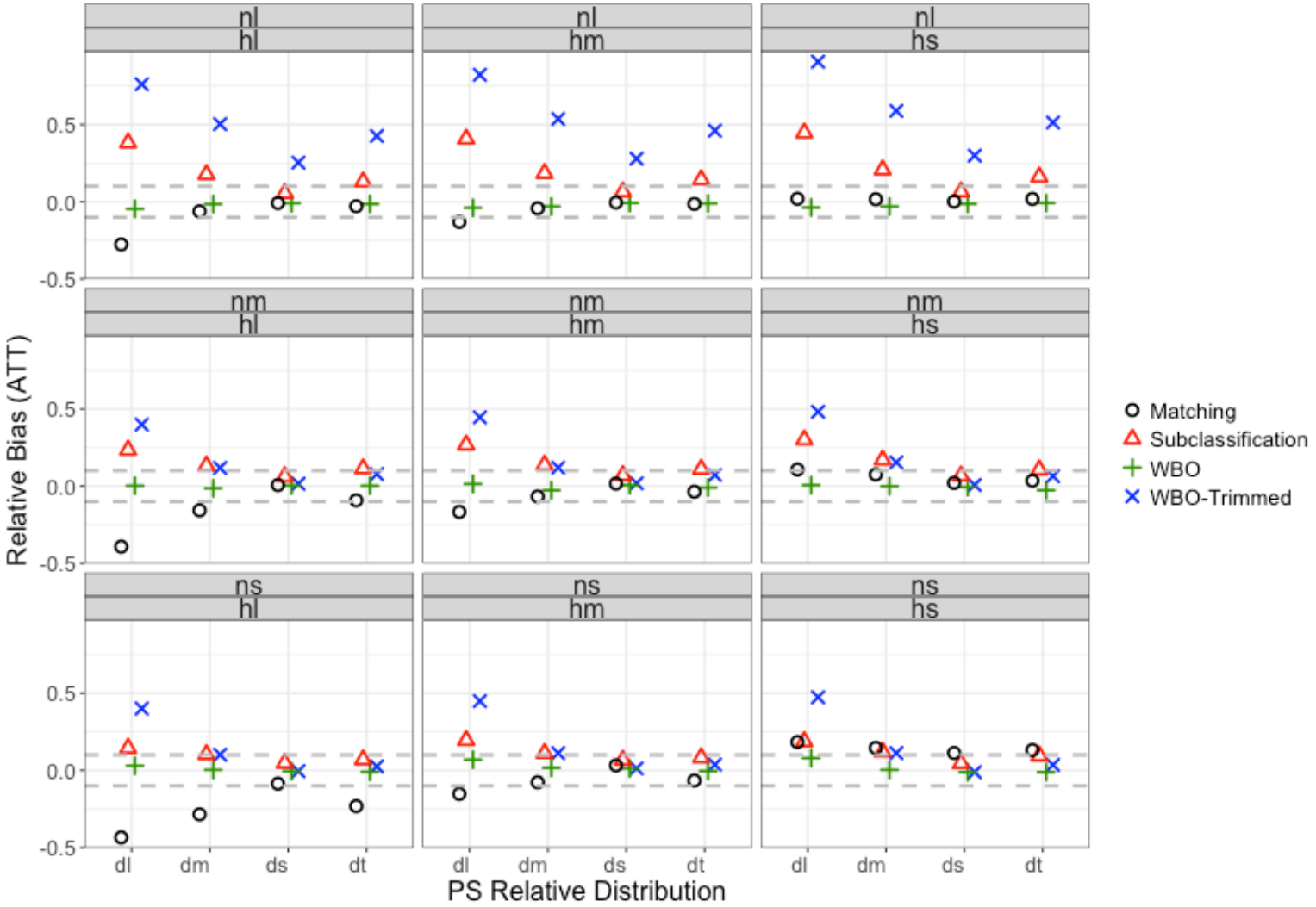


Figure 6. Relative bias for ATT methods (Scenario A). The plot for scenario B is presented in Appendix B. dl(dm/ds/dt) = large(medium/small/truncated) difference in relative PSdistributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects, another way of saying substantial, slight, and zero heterogeneous treatment effects discussed in Chapter 3.

Table 5

Relative Bias in ATT Estimate

Conditions			Relative Bias (ATT)				
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Matching	Subclassification	WBO	WBO-Trimmed
Large	Truncated	Substantial	6.61	-0.03	0.13	-0.01	0.43
Large	Truncated	Slight	2.28	-0.01	0.14	-0.01	0.46
Large	Truncated	Zero	1.03	0.02	0.16	-0.01	0.51
Large	Large	Substantial	-7.81	-0.28	0.38	-0.05	0.76
Large	Large	Slight	9.14	-0.13	0.41	-0.04	0.82
Large	Large	Zero	1.75	0.02	0.45	-0.04	0.91
Large	Medium	Substantial	20.63	-0.06	0.18	-0.02	0.50
Large	Medium	Slight	3.15	-0.04	0.18	-0.03	0.54
Large	Medium	Zero	1.20	0.02	0.21	-0.03	0.59
Large	Small	Substantial	1.78	-0.01	0.06	-0.01	0.25
Large	Small	Slight	1.08	-0.01	0.06	-0.01	0.28
Large	Small	Zero	0.61	0.00	0.06	-0.01	0.30
Medium	Truncated	Substantial	4.02	-0.09	0.11	0.00	0.08
Medium	Truncated	Slight	1.93	-0.04	0.11	-0.01	0.07
Medium	Truncated	Zero	1.01	0.03	0.11	-0.03	0.06
Medium	Large	Substantial	-21.11	-0.39	0.23	0.00	0.40
Medium	Large	Slight	5.87	-0.17	0.27	0.01	0.45
Medium	Large	Zero	1.81	0.11	0.30	0.01	0.48
Medium	Medium	Substantial	7.13	-0.16	0.13	-0.01	0.12
Medium	Medium	Slight	2.57	-0.07	0.14	-0.03	0.12
Medium	Medium	Zero	1.23	0.07	0.17	0.00	0.15
Medium	Small	Substantial	1.43	0.01	0.06	0.00	0.01
Medium	Small	Slight	0.97	0.02	0.07	0.01	0.02
Medium	Small	Zero	0.62	0.02	0.07	-0.01	0.01
Small	Truncated	Substantial	2.16	-0.23	0.07	-0.01	0.03
Small	Truncated	Slight	1.51	-0.07	0.08	0.00	0.04
Small	Truncated	Zero	1.03	0.13	0.10	-0.01	0.04
Small	Large	Substantial	9.16	-0.43	0.14	0.03	0.40
Small	Large	Slight	3.53	-0.15	0.20	0.07	0.45
Small	Large	Zero	1.84	0.18	0.19	0.08	0.47
Small	Medium	Substantial	3.09	-0.28	0.10	0.00	0.10
Small	Medium	Slight	1.92	-0.08	0.11	0.02	0.11
Small	Medium	Zero	1.22	0.15	0.12	0.00	0.11
Small	Small	Substantial	1.00	-0.09	0.05	0.00	0.00
Small	Small	Slight	0.83	0.03	0.07	0.01	0.01
Small	Small	Zero	0.60	0.11	0.05	-0.01	-0.01

Note. The corresponding table for Scenario B is presented in Appendix E.

Table 6

PS Matching Rate

Scenario	Sample Size	Rel. PS Dist. L			Rel. PS Dis. M			Rel. PS Dist. S			Rel. PS Dist. T		
		Heter. TE			Heter. TE			Heter. TE			Heter. TE		
		L	M	S	L	M	S	L	M	S	L	M	S
A	L	0.76	0.77	0.76	0.95	0.95	0.95	1.00	1.00	1.00	0.98	0.98	0.98
	M	0.55	0.55	0.56	0.81	0.81	0.82	0.98	0.98	0.98	0.89	0.89	0.89
	S	0.34	0.34	0.34	0.54	0.55	0.55	0.79	0.78	0.79	0.62	0.62	0.62
B	L	0.77	0.77	0.77	0.95	0.95	0.95	1.00	1.00	1.00	0.98	0.98	0.98
	M	0.55	0.56	0.55	0.82	0.82	0.81	0.98	0.98	0.98	0.89	0.89	0.89
	S	0.35	0.35	0.35	0.54	0.55	0.55	0.78	0.78	0.78	0.62	0.62	0.62

Note. L = large, M = medium, S = small, and T = truncated. The L, M, S for TE heterogeneity are the same as substantial, slight, and zero respectively.

With larger sample sizes in the control group, although the *proportion* of cases with high propensity scores in the control group was still relatively low compared to that for lower propensity scores, the actual *number* of cases with high propensity scores was bigger such that the pool to find a match (i.e., without replacement) from the control group for the treated cases was larger than the conditions with smaller sample sizes in the control group. Therefore, a larger sample size in the control group mitigated the bias by increasing the matching rate between the two groups. In particular, the correlation between the matching rate and the absolute values of relative bias was -0.73 across all conditions. The negative relation between matching rate and relative bias is shown in Figure 7.

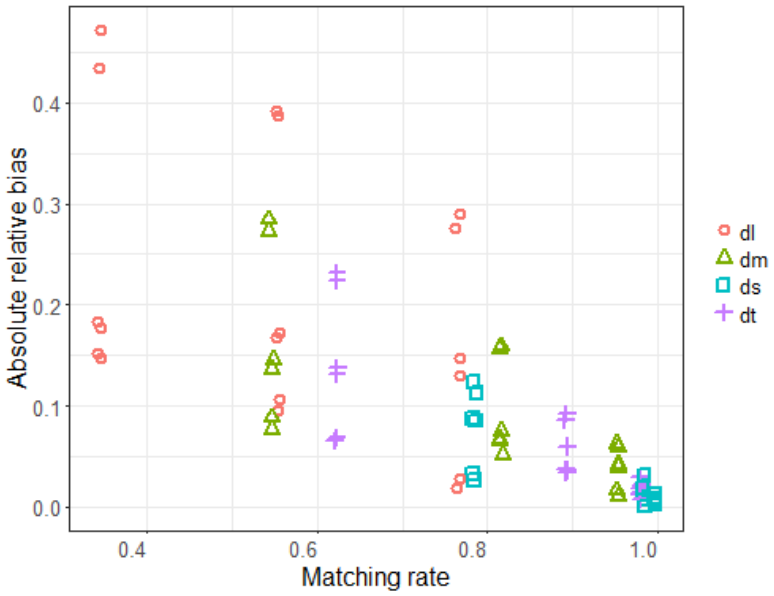


Figure 7. The correlation between matching rate and absolute relative bias. Each dot represents one of the 72 conditions (e.g., large relative difference in PS distribution–small sample size in the control group–large heterogeneous treatment effects–Scenario A). Only different levels of PS relative distributions are differentiated in the chart.

Interactions between the *heterogeneous treatment effects* and the other factors (i.e., PS distribution and sample size) were also observed in the resulting bias in the TE estimates. Specifically, in the way the data were simulated for this study, the subjects that did not have matches from the control group and thus were removed from the treatment group tended to have higher propensity scores with large negative individual treatment effects (the levels of propensity scores and the associated individual treatment effects are shown in Table 7); as a result, the more subjects removed, the more bias was produced (positive bias in particular) in the TE estimates. This is also the reason why the relative bias for PS matching tended to be negative (i.e., positive bias divided by negative true treatment effect). When the treatment effects were homogeneous, the results were robust to the PS relative distributions. This is because when all subjects have the

same individual treatment effects regardless of the associated propensity scores, the remaining cases are always representative of the treatment population, such that *how many* and *which* cases were removed do not affect the accuracy of the TE estimates.

Table 7

Empirical Average Individual Treatment Effects at Different Levels of Propensity Scores

PS Level	Heter. TE (L)		Heter. TE (M)		Heter. TE (S)	
	ITE (ATT)	ITE (ATE)	ITE (ATT)	ITE (ATE)	ITE (ATT)	ITE (ATE)
0 ~ 0.1	0.40	0.57	0.00	0.08	-0.40	-0.40
0.1 ~ 0.2	0.18	0.18	-0.10	-0.10	-0.40	-0.40
0.2 ~ 0.3	0.06	0.06	-0.15	-0.16	-0.40	-0.40
0.3 ~ 0.4	0.00	0.00	-0.21	-0.20	-0.40	-0.40
0.4 ~ 0.5	-0.09	-0.08	-0.24	-0.24	-0.40	-0.40
0.5 ~ 0.6	-0.15	-0.15	-0.29	-0.29	-0.40	-0.40
0.6 ~ 0.7	-0.22	-0.22	-0.32	-0.31	-0.40	-0.40
0.7 ~ 0.8	-0.31	-0.31	-0.36	-0.36	-0.40	-0.40
0.8 ~ 0.9	-0.42	-0.42	-0.42	-0.42	-0.40	-0.40
0.9 ~ 1	-0.81	-0.80	-0.61	-0.60	-0.40	-0.40
overall	-.048	-0.12	-0.44	-0.26	-0.40	-0.40

Note. The numbers are from Scenario A, for the conditions of large relative PS distribution and small sample size in the control group. L, M, and S refer to substantial, slight, and zero heterogeneity in treatment effects. The averaged individual treatment effects (ITE) for ATT were calculated based on the treatment group only; the averaged ITEs for the ATE were averaged across both groups.

4.1.3 Precision

The precision for TE estimates (i.e., the empirical SE for each condition was computed as the standard deviation of the TE estimates across all 250 replications) generated by PS matching was sensitive to the PS relative distributions when there was a small sample size in the control group; specifically, better precision was observed when the differences in the relative PS distributions were small. With medium and large sample sizes in the control group, the precision of the TE estimation was relatively robust to relative PS distributions (Figure 8). The empirical SEs were the lowest with a large sample size in the control group, also possibly due to a higher matching rate (Table 6, Table 8). Compared to the other PS conditioning methods, PS matching

demonstrated the best precision in the condition with a small sample size in the control group. The precision for PS matching did not vary across different levels of heterogeneity of treatment effects.

Table 8

Empirical Standard Error for ATT Estimate

Conditions			Empirical Standard Error (ATT)				
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Matching	Subclassification	WBO	WBO-Trimmed
Large	Truncated	Substantial	0.06	0.06	0.05	0.05	0.05
Large	Truncated	Slight	0.04	0.05	0.04	0.04	0.04
Large	Truncated	Zero	0.04	0.05	0.03	0.04	0.04
Large	Large	Substantial	0.06	0.06	0.06	0.08	0.06
Large	Large	Slight	0.05	0.05	0.05	0.07	0.05
Large	Large	Zero	0.04	0.05	0.04	0.07	0.04
Large	Medium	Substantial	0.05	0.06	0.05	0.05	0.05
Large	Medium	Slight	0.04	0.05	0.04	0.04	0.04
Large	Medium	Zero	0.04	0.05	0.04	0.04	0.04
Large	Small	Substantial	0.06	0.06	0.05	0.05	0.05
Large	Small	Slight	0.05	0.05	0.04	0.04	0.04
Large	Small	Zero	0.04	0.05	0.03	0.03	0.04
Medium	Truncated	Substantial	0.06	0.05	0.05	0.06	0.06
Medium	Truncated	Slight	0.05	0.05	0.04	0.05	0.05
Medium	Truncated	Zero	0.05	0.05	0.04	0.05	0.05
Medium	Large	Substantial	0.06	0.07	0.07	0.12	0.07
Medium	Large	Slight	0.05	0.07	0.07	0.11	0.07
Medium	Large	Zero	0.05	0.06	0.06	0.11	0.07
Medium	Medium	Substantial	0.06	0.06	0.06	0.07	0.07
Medium	Medium	Slight	0.05	0.06	0.05	0.07	0.06
Medium	Medium	Zero	0.04	0.05	0.04	0.05	0.05
Medium	Small	Substantial	0.06	0.06	0.05	0.06	0.06
Medium	Small	Slight	0.05	0.05	0.04	0.04	0.04
Medium	Small	Zero	0.05	0.05	0.04	0.04	0.04
Small	Truncated	Substantial	0.07	0.07	0.07	0.08	0.07
Small	Truncated	Slight	0.06	0.06	0.07	0.08	0.08
Small	Truncated	Zero	0.06	0.05	0.07	0.08	0.07
Small	Large	Substantial	0.07	0.09	0.15	0.20	0.10
Small	Large	Slight	0.06	0.08	0.16	0.21	0.10
Small	Large	Zero	0.05	0.08	0.14	0.19	0.09
Small	Medium	Substantial	0.07	0.07	0.08	0.10	0.08
Small	Medium	Slight	0.06	0.06	0.07	0.09	0.07
Small	Medium	Zero	0.06	0.06	0.07	0.09	0.07
Small	Small	Substantial	0.07	0.06	0.06	0.07	0.07
Small	Small	Slight	0.06	0.05	0.05	0.05	0.05
Small	Small	Zero	0.05	0.05	0.05	0.05	0.05

Note. The corresponding table for Scenario B is presented in Appendix E.

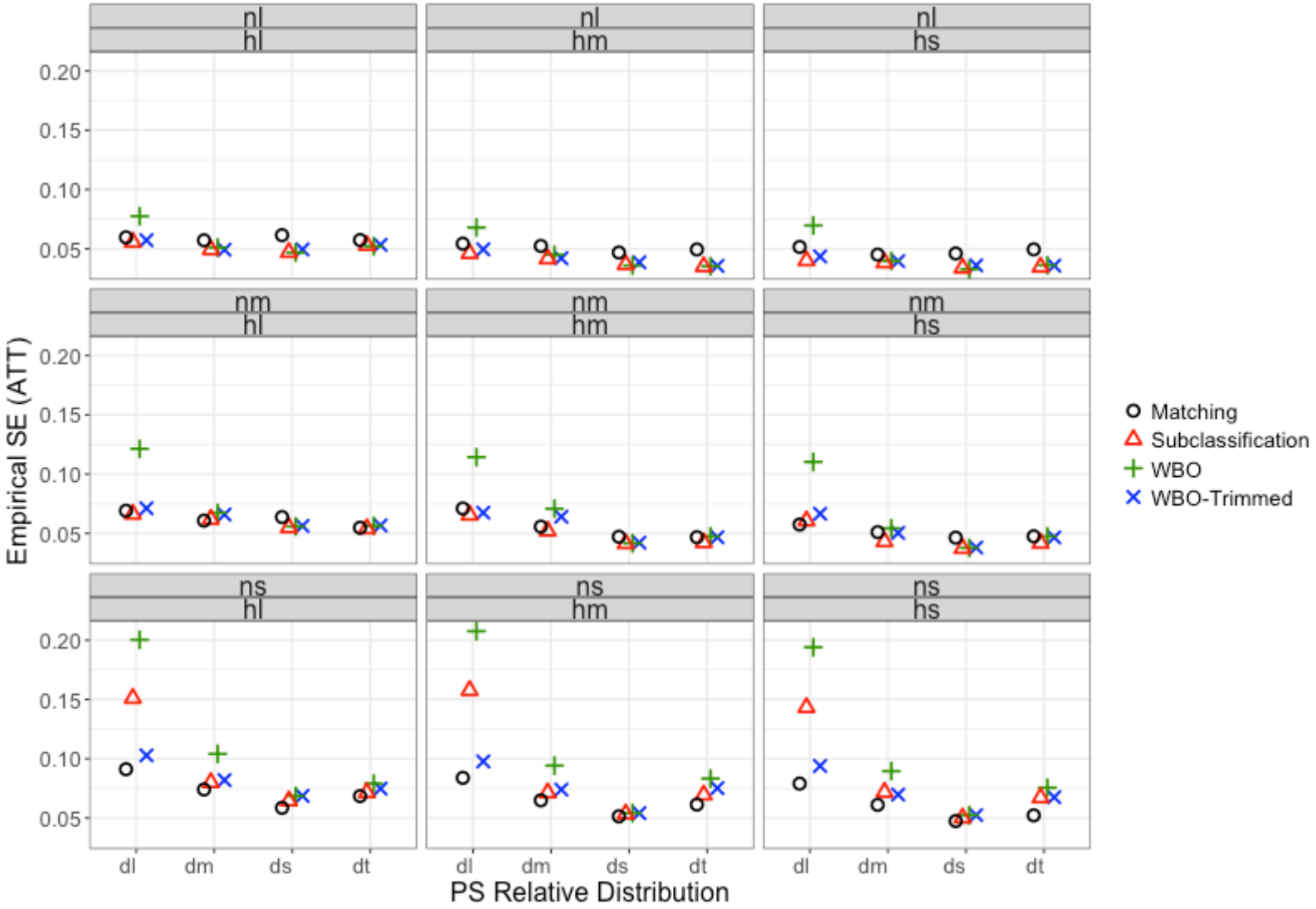


Figure 8. Standard Error for ATT methods (Scenario A). The plot for scenario B is presented in Appendix C. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects.

4.1.4 Relative bias in SE estimate

Having demonstrated relatively low and consistent precision across different data conditions compared to other PS conditioning methods, PS matching is not necessarily always the best method that yields the most trustworthy SE estimates. As is shown in Figure 9 and Table 9, although the *fluctuation* of the relative bias in SE estimates was relatively small compared to subclassification and even WBO, there were only 3 conditions when PS matching eventually produced *trustworthy* SE estimates (< 10% according to Table 4) and 4 conditions when it could be used *with caution* (>= 10% & < 15%). When the sample size was small, the SE estimates never met the thresholds. The bias is also found at the sampling variance level, so the relative bias in SE estimates was not just the result of non-linear transformation of the variance to the SE. This conclusion applies to the results of the other PS conditioning methods.

Table 9

Relative Bias in the SE Estimate for ATT

Conditions			Relative Bias in SE Estimate (ATT)				
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Matching	Subclassification	WBO	WBO-Trimmed
Large	Truncated	Substantial	-0.23	0.16	0.58	0.12	0.07
Large	Truncated	Slight	0.05	0.20	1.32	0.40	0.34
Large	Truncated	Zero	0.05	0.15	1.33	0.29	0.24
Large	Large	Substantial	-0.20	0.20	0.33	0.07	0.13
Large	Large	Slight	-0.05	0.17	0.52	0.08	0.14
Large	Large	Zero	0.17	0.20	0.72	0.02	0.23
Large	Medium	Substantial	-0.13	0.17	0.65	0.19	0.18
Large	Medium	Slight	-0.01	0.14	0.89	0.17	0.18
Large	Medium	Zero	-0.02	0.27	1.04	0.25	0.17
Large	Small	Substantial	-0.20	0.10	0.86	0.21	0.13
Large	Small	Slight	-0.05	0.28	1.33	0.32	0.21
Large	Small	Zero	-0.02	0.25	1.52	0.33	0.20
Medium	Truncated	Substantial	-0.11	0.26	0.94	0.22	0.20
Medium	Truncated	Slight	0.06	0.32	1.32	0.30	0.30
Medium	Truncated	Zero	0.00	0.25	1.34	0.24	0.25
Medium	Large	Substantial	-0.08	0.24	0.84	-0.08	0.17
Medium	Large	Slight	-0.03	0.08	0.66	-0.06	0.16
Medium	Large	Zero	0.00	0.28	0.74	-0.07	0.14
Medium	Medium	Substantial	-0.18	0.18	0.72	0.13	0.10
Medium	Medium	Slight	-0.07	0.16	0.91	0.00	0.03
Medium	Medium	Zero	0.07	0.22	1.28	0.22	0.25
Medium	Small	Substantial	-0.18	0.06	0.88	0.10	0.09
Medium	Small	Slight	-0.03	0.28	1.36	0.29	0.27
Medium	Small	Zero	-0.05	0.24	1.60	0.33	0.32
Small	Truncated	Substantial	-0.01	0.22	1.19	0.16	0.16
Small	Truncated	Slight	0.01	0.22	1.02	0.05	0.07
Small	Truncated	Zero	0.00	0.38	1.07	0.13	0.18
Small	Large	Substantial	-0.03	0.25	0.78	-0.21	0.04
Small	Large	Slight	0.04	0.21	0.48	-0.25	0.05
Small	Large	Zero	0.03	0.22	0.59	-0.22	0.08
Small	Medium	Substantial	-0.03	0.21	1.16	0.02	0.12
Small	Medium	Slight	-0.01	0.22	1.13	0.05	0.15
Small	Medium	Zero	-0.05	0.25	1.09	0.11	0.21
Small	Small	Substantial	0.04	0.28	1.17	0.09	0.09
Small	Small	Slight	0.00	0.31	1.38	0.27	0.27
Small	Small	Zero	0.06	0.36	1.51	0.26	0.26

Note. The corresponding table for Scenario B is presented in Appendix E.

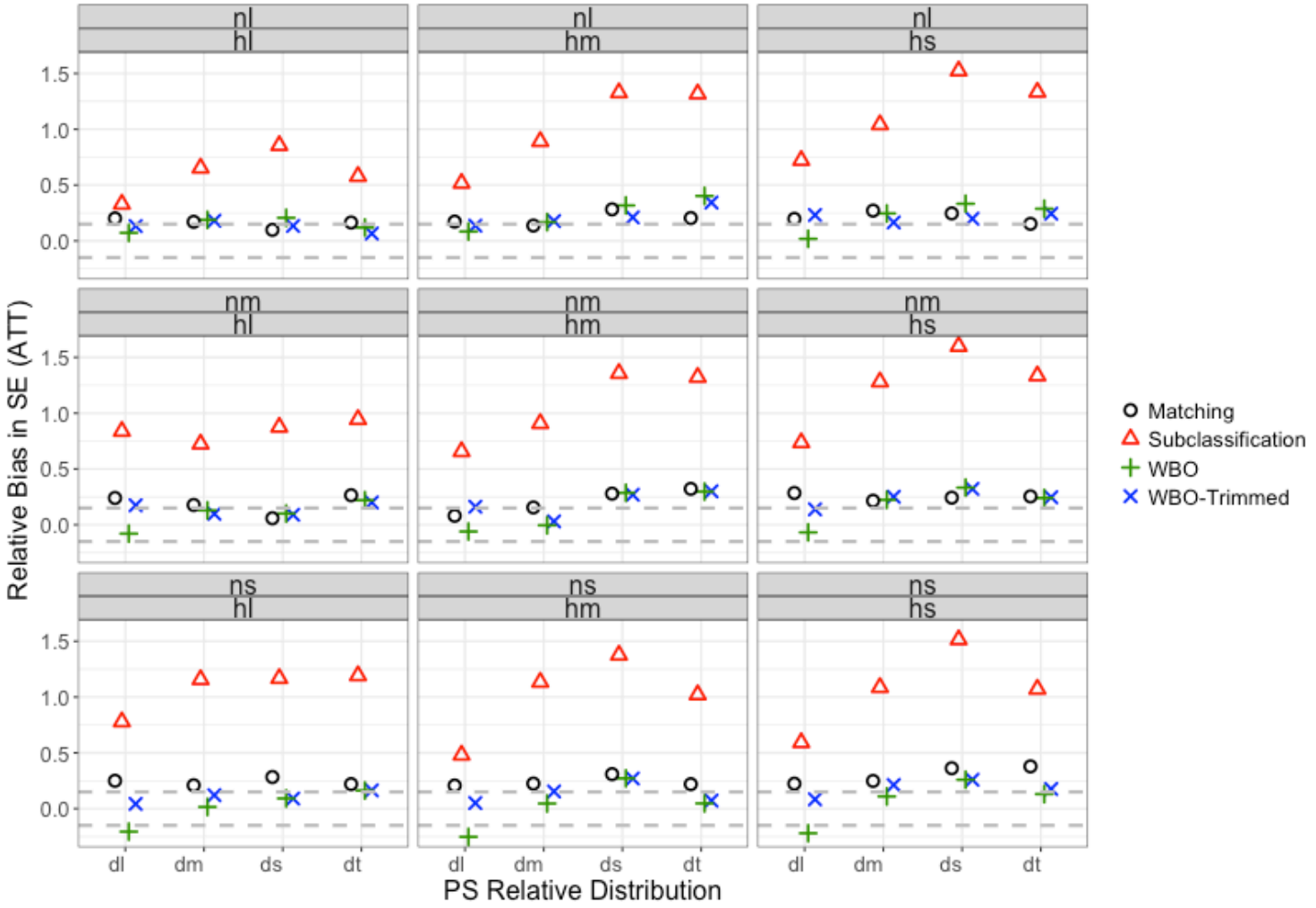


Figure 9. Relative bias in SE for ATT methods (Scenario A). The plot for scenario B is presented in Appendix D. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects. The grey dotted lines indicate the thresholds between *use with caution* and *untrustworthy* (±15%).

4.2 Subclassification

PS subclassification was implemented to estimate both the ATT and the ATE in the simulation study. As a reminder, subclassification was conducted using 5 classes defined by an even split of ALL subjects ordered by PS estimates into classes. For the subclasses that had fewer than 2 cases, they were collapsed with the neighbor subclass. This section analyzes the results for both the ATT and ATE estimates, in terms of covariate balance, estimation accuracy, precision, and how trustworthy the SE estimate was (i.e., relative bias in SE estimates).

4.2.1 Balance

The covariate balance using subclassfication was not as good as that found with PS matching. Although most of the balance values fell under the relaxed criterion of 0.25, they were not acceptable in the conditions with large PS relative distributions with medium and large sample sizes in the control group. This is because the subclasses were generated based on the quintiles of the *overall* sample, and the subclasses associated with the lowest propensity scores would have very few cases in the treated group, in the conditions indicated above. That said, the balance, and therefore the TE estimates for those subclasses (as will be discussed in Sections 4.2.2 and 4.2.3), would be dependent on sampling fluctuation and might not reflect the true relations in the population. This was more of an issue for the ATE than the ATT, because the results from these imbalanced subclasses were weighted minimally with the ATT, however, were weighted equally as the other subclasses with ATE. The fact that the balance for these conditions was bad indicated the fact that the treated and control groups were not comparable, and therefore the TE estimates obtained in these conditions may not be accurate.

4.2.2 Accuracy of ATT estimation

With smaller differences in the distributions across treat and control, PS subclassification, like PS matching, demonstrated better accuracy in ATT estimates. At the same time, PS subclassification produced more bias under the condition of a larger sample size in the control group and it did not show a clear difference in performance across the conditions of TE heterogeneity when holding the other conditions constant.

The way the data were simulated resulted in more subjects clustering at higher propensity scores in the treatment group and at lower propensity scores in the control group. Subclassified by the quintiles of the propensity scores in the *overall* sample (i.e., both control and treat), more treated cases fell in the subclass with higher propensity scores. Because the ATT is weighted by the number of cases in the treated group, the highest subclass thus weighted most heavily in the calculation of the ATT. What is more, because higher propensity scores were associated with lower individual treatment effects (shown in Table 7) in the heterogeneous treatment effect conditions, when subclasses with higher propensity scores received more weight, the TE estimates tended to be negatively biased. The relative bias shown in Table 5 and Figure 6 were thus all positive for the reason that both bias and true treatment effect were negative. Under the condition of a larger difference in PS relative distributions, even more treated subjects relative to controlled ones would fall in the higher-propensity-score subclasses, and therefore generated more bias in general.

With respect to sample size, when the sample size is very big in the control group, the PS distribution of the control group dominates the subclassification. In particular, when subclassifying by the quintiles of the propensity scores, almost all subjects in the subclasses associated with lower propensity scores are from the control group while very few are in the

treatment group. As a result, in the calculation of the overall ATT estimate, the subclasses with lower propensity scores take little weight compared to those with higher propensity scores. This issue, similar to that found in the condition with the large difference in PS relative distributions, results in greater negative bias and accordingly greater positive relative bias due to negative true treatment effects.

The relative bias in ATT estimates produced by subclassification did not seem to change across different levels of TE heterogeneity. Because the propensity scores and ITEs were simulated to be negatively correlated, larger heterogeneous treatment effects resulted in more extreme individual treatment effects at both ends of propensity scores – greater negative ITEs associated with higher propensity scores and positive ITEs associated with lower propensity scores. With more weight received by subclasses with higher propensity scores, as indicated in the previous paragraphs, the overall ATT estimates in the conditions with greater heterogeneous treatment effects tended to decrease (i.e., negatively greater, shown in Figure 10).

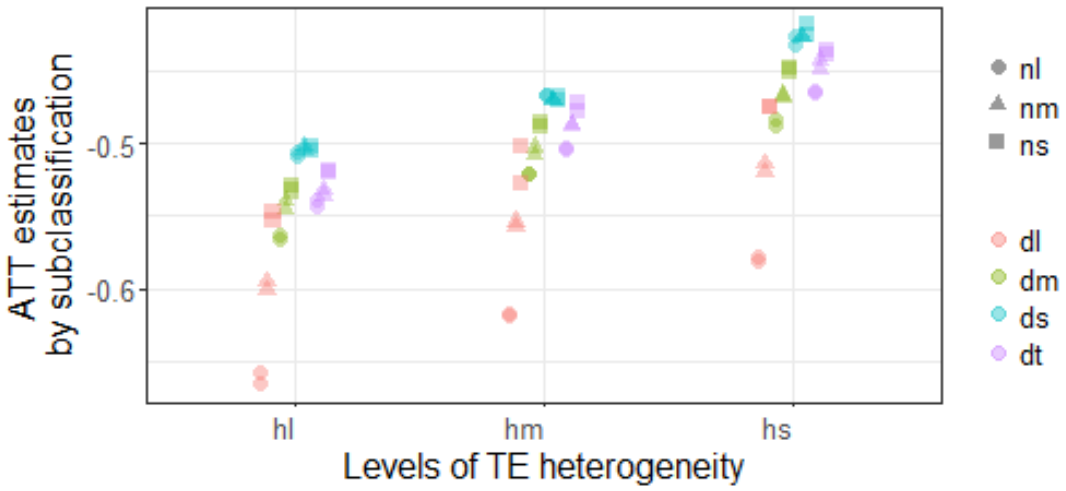


Figure 10. ATT estimates at different levels of TE heterogeneity. The two dots for each shape (conditions of sample size) and color (conditions PS relative distributions) respectively represent the two scenarios (A and B) with correct and incorrect PS specifications).

Given that the ATT estimates varied at different levels of TE heterogeneity, why was the relative bias robust to TE heterogeneity? This is because the relative bias is determined by both ATT estimates and true ATTs (Equation 22). In this simulation study, in addition to the changes observed in ATT estimates, the true treatment effects also decrease (i.e., increase negatively) as the level of heterogeneity goes up (true ATT distributions shown in Figure 11). This is the way the data were generated. Specifically, there is a statistically significantly positive linear relation between the true ATT values and the ATT estimates ($r = .61$, $p < .001$, shown in the left panel of Figure 12). The relative bias, accordingly, are relatively constant across different levels of TE heterogeneity (right panel of Figure 12).

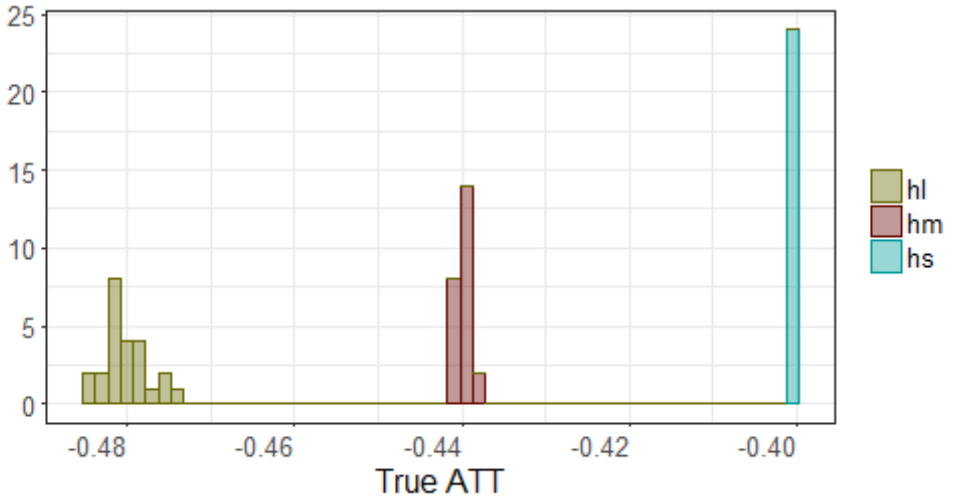


Figure 11. The distributions of true ATTs at different levels of TE heterogeneity. The means of the ATTs for zero, medium, and high levels of TE heterogeneity are -.40, -.44, and -.48 respectively.

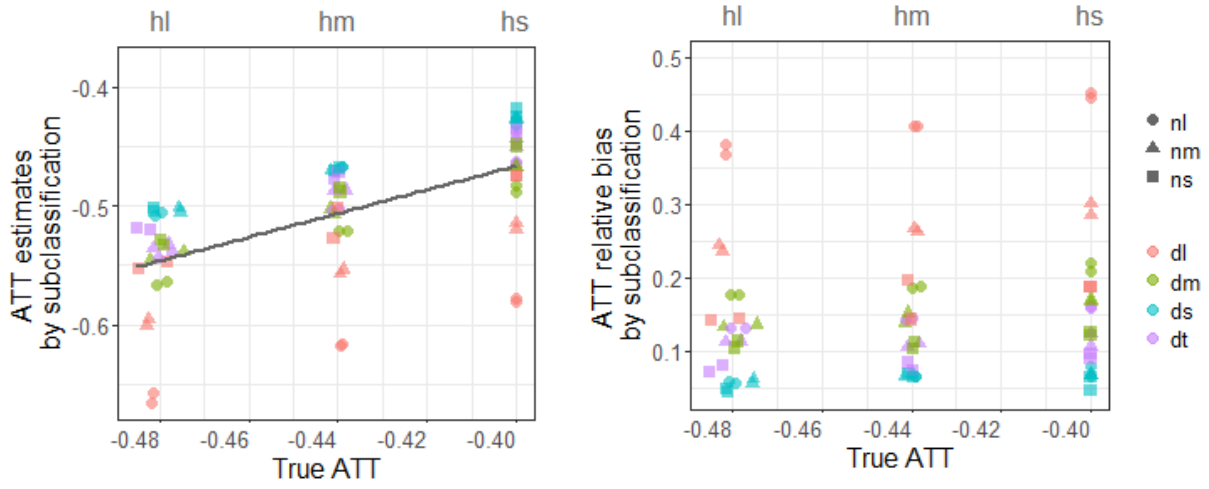


Figure 12. The relation between true ATTs and subclassification ATT estimates (left) and the relation between true ATT and subclassification relative bias (right).

4.2.3 Accuracy of ATE estimation

For ATE estimation with PS subclassification, it is still the case that larger bias was associated with greater difference in PS relative distributions, as well as larger sample size in the

control group, except that in some extreme cases (i.e., the combination of larger size of control and difference in PS relative distributions) the relative bias was negative (Table 10, Figure 13). Unlike subclassification for ATT, subclassification for ATE was *seemingly* sensitive to heterogeneous treatment effects, with large differene in PS relative distributions and larger sample size in the control group. However, this was more of the issue of imbalanced samples between the two groups. As indicated in Section 4.2.1, in these conditions (i.e., medium and large sample sizes in the control group with large PS relative distribution), the high-propensity-score subclass only included few observations which may not be sufficient to produce reasonable balance and to represent the true treatment effects. When these subclasses were weighted equally as the other subclasses in producing the final ATE estimate, it introduced more bias.

Table 10

Relative Bias in ATE Estimate

Conditions			Relative Bias (ATE)			
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Subclassification	IPTW	IPTW-Trimmed
Large	Truncated	Substantial	6.61	0.72	0.09	3.54
Large	Truncated	Slight	2.28	0.21	-0.03	1.19
Large	Truncated	Zero	1.03	0.09	-0.03	0.52
Large	Large	Substantial	-7.81	-1.30	-0.26	-4.43
Large	Large	Slight	9.14	1.38	0.14	5.10
Large	Large	Zero	1.75	0.17	-0.01	0.95
Large	Medium	Substantial	20.63	2.18	0.07	11.06
Large	Medium	Slight	3.15	0.30	-0.06	1.66
Large	Medium	Zero	1.20	0.08	-0.05	0.61
Large	Small	Substantial	1.78	0.17	-0.02	0.96
Large	Small	Slight	1.08	0.10	-0.01	0.57
Large	Small	Zero	0.61	0.06	0.00	0.32
Medium	Truncated	Substantial	4.02	0.41	0.02	0.86
Medium	Truncated	Slight	1.93	0.18	-0.01	0.37
Medium	Truncated	Zero	1.01	0.09	-0.01	0.18
Medium	Large	Substantial	-21.11	-2.73	-0.72	-9.23
Medium	Large	Slight	5.87	0.71	0.09	2.51
Medium	Large	Zero	1.81	0.21	0.04	0.73
Medium	Medium	Substantial	7.13	0.69	0.02	1.98
Medium	Medium	Slight	2.57	0.25	-0.01	0.67
Medium	Medium	Zero	1.23	0.12	0.02	0.30
Medium	Small	Substantial	1.43	0.16	0.02	0.12
Medium	Small	Slight	0.97	0.10	-0.01	0.06
Medium	Small	Zero	0.62	0.06	0.00	0.03
Small	Truncated	Substantial	2.16	0.20	-0.02	0.11
Small	Truncated	Slight	1.51	0.15	0.01	0.09
Small	Truncated	Zero	1.03	0.09	-0.01	0.05
Small	Large	Substantial	9.16	0.88	0.12	3.32
Small	Large	Slight	3.53	0.39	0.14	1.27
Small	Large	Zero	1.84	0.17	0.06	0.65
Small	Medium	Substantial	3.09	0.32	-0.01	0.39
Small	Medium	Slight	1.92	0.21	0.02	0.25
Small	Medium	Zero	1.22	0.12	0.00	0.15
Small	Small	Substantial	1.00	0.09	-0.01	-0.01
Small	Small	Slight	0.83	0.09	0.01	0.01
Small	Small	Zero	0.60	0.05	-0.02	-0.01

Note. The corresponding table for Scenario B is presented in Appendix E.

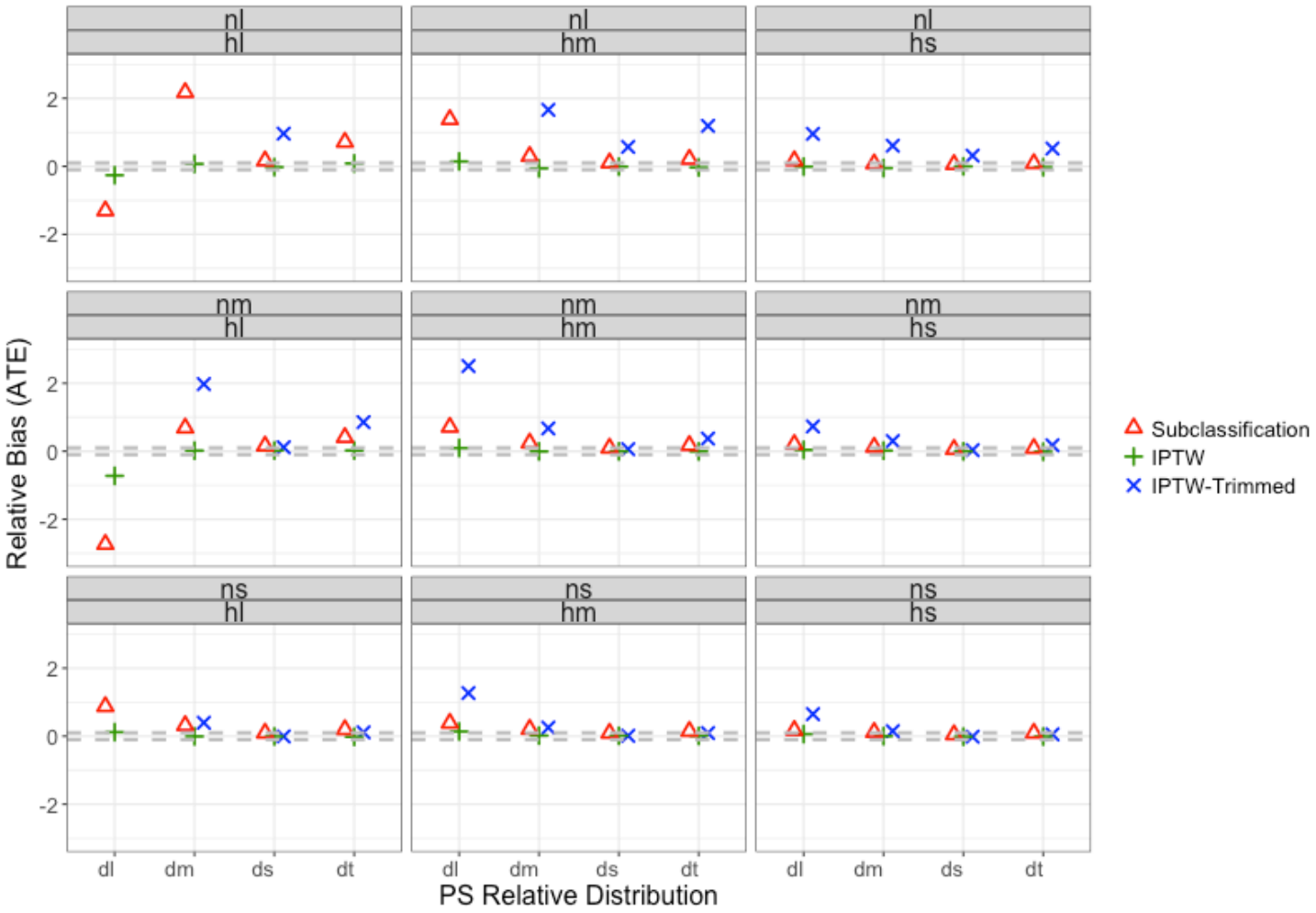


Figure 13. Relative bias for ATE methods (Scenario A). dl (dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects. Five large relative bias produced by IPTW with trimming are out of boundary of the plot. The values are -4.43, 11.06, 3.54, 5.10, -9.23, and 3.31 for nl_dl_hl, nl_dm_hl, nl_dt_hl, hl_dl_hm, nm_dl_hl, ns_dl_hl, respectively. The plot for scenario B is presented in Appendix B.

4.2.4 Precision

With subclassification, the standard errors for ATT estimates were sensitive to PS relative distributions only with small sample size in the control group (see Figure 8, Table 8); for ATE estimation, they were sensitive to PS relative distributions across all sample size conditions (see Figure 14, Table 11). Similar to PS matching, the precision was not affected by different levels of heterogeneity of treatment effects. As was discussed in Chapter 4, this metric is good to keep in mind however was not used to measure the performance of the models, because the SE estimates may not even reflect the true SE of an estimation in real life. Therefore, the performance of the relative bias in SE estimate will be discussed in the next section (Section 4.2.5).

Table 11

Empirical Standard Error for ATE Estimate

Conditions			Empirical Standard Error (ATE)			
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Subclassification	IPTW	IPTW-Trimmed
Large	Truncated	Substantial	0.06	0.07	0.09	0.05
Large	Truncated	Slight	0.04	0.06	0.09	0.04
Large	Truncated	Zero	0.04	0.05	0.07	0.04
Large	Large	Substantial	0.06	0.16	0.29	0.06
Large	Large	Slight	0.05	0.12	0.23	0.05
Large	Large	Zero	0.04	0.12	0.19	0.04
Large	Medium	Substantial	0.05	0.08	0.13	0.05
Large	Medium	Slight	0.04	0.07	0.11	0.04
Large	Medium	Zero	0.04	0.07	0.09	0.04
Large	Small	Substantial	0.06	0.05	0.06	0.05
Large	Small	Slight	0.05	0.04	0.04	0.04
Large	Small	Zero	0.04	0.04	0.04	0.04
Medium	Truncated	Substantial	0.06	0.06	0.08	0.05
Medium	Truncated	Slight	0.05	0.05	0.06	0.04
Medium	Truncated	Zero	0.05	0.05	0.06	0.04
Medium	Large	Substantial	0.06	0.14	0.26	0.07
Medium	Large	Slight	0.05	0.11	0.26	0.06
Medium	Large	Zero	0.05	0.11	0.18	0.05
Medium	Medium	Substantial	0.06	0.08	0.12	0.06
Medium	Medium	Slight	0.05	0.06	0.09	0.05
Medium	Medium	Zero	0.04	0.06	0.07	0.05
Medium	Small	Substantial	0.06	0.05	0.05	0.05
Medium	Small	Slight	0.05	0.04	0.04	0.04
Medium	Small	Zero	0.05	0.04	0.04	0.04
Small	Truncated	Substantial	0.07	0.06	0.08	0.07
Small	Truncated	Slight	0.06	0.06	0.07	0.06
Small	Truncated	Zero	0.06	0.05	0.06	0.05
Small	Large	Substantial	0.07	0.13	0.25	0.09
Small	Large	Slight	0.06	0.12	0.19	0.07
Small	Large	Zero	0.05	0.12	0.19	0.07
Small	Medium	Substantial	0.07	0.07	0.11	0.07
Small	Medium	Slight	0.06	0.06	0.09	0.06
Small	Medium	Zero	0.06	0.06	0.07	0.06
Small	Small	Substantial	0.07	0.06	0.05	0.05
Small	Small	Slight	0.06	0.05	0.05	0.05
Small	Small	Zero	0.05	0.04	0.04	0.04

Note. The corresponding table for Scenario B is presented in Appendix E.

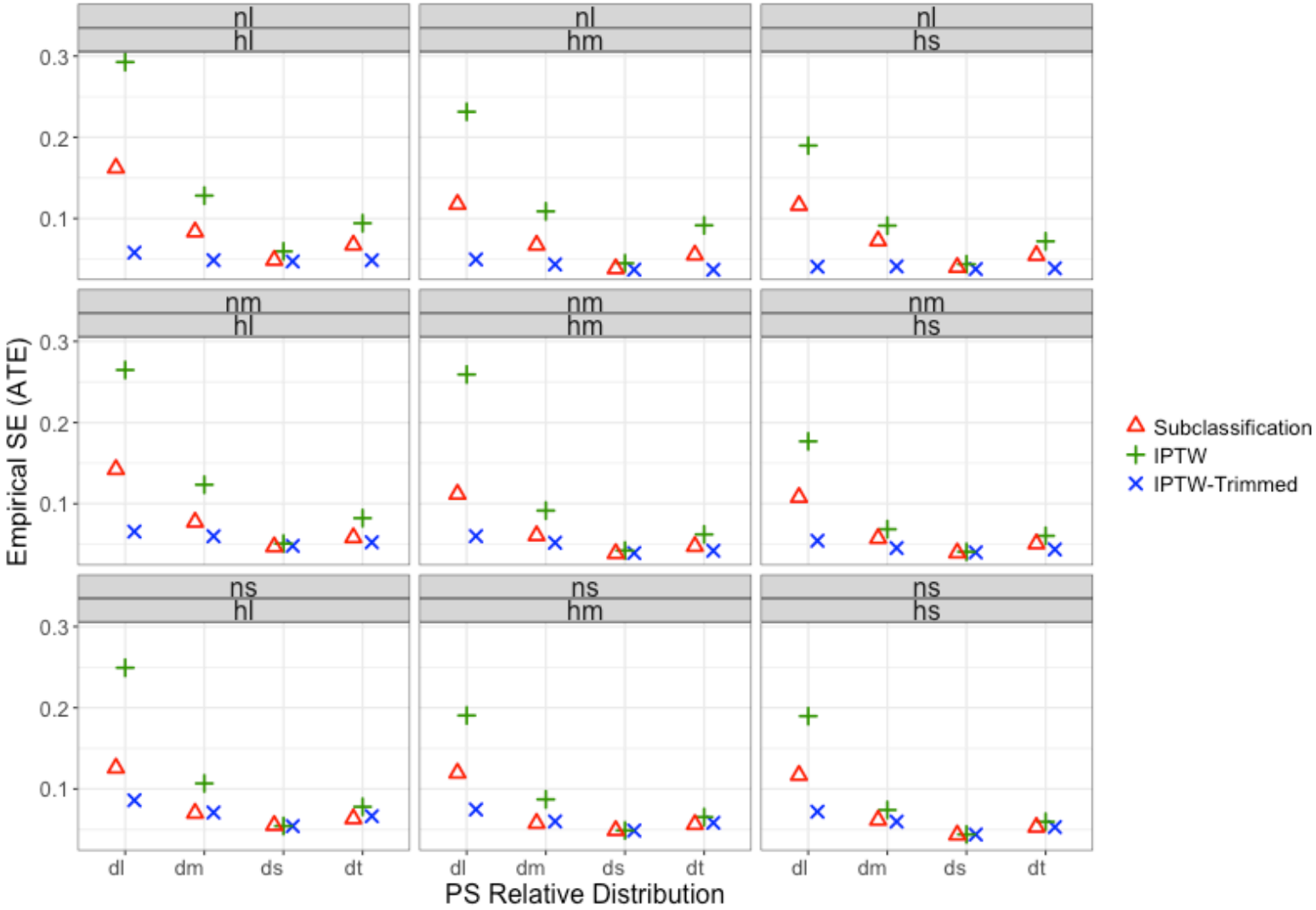


Figure 14. Standard Error for ATE methods (Scenario A). The plot for scenario B is presented in Appendix C. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects.

4.2.5 Relative Bias in SE Estimate

Although the empirical SEs for treatment effects associated with subclassification methods can be low (meaning better precision) relative to other PS conditioning methods, especially when the PS relative distribution was small, the relative bias in SE was the largest among all PS conditioning methods (Table 9, Table 12, Figure 9, Figure 15). According to guidelines for results thresholds shown in Table 4, the SE estimates were never trustworthy, with relative bias in SE always >50%.

Table 12

Relative Bias in the SE Estimate for ATE

Conditions			Relative Bias in SE Estimate (ATE)			
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Subclassification	IPTW	IPTW-Trimmed
Large	Truncated	Substantial	-0.23	0.87	0.08	0.16
Large	Truncated	Slight	0.05	1.25	-0.07	0.30
Large	Truncated	Zero	0.05	1.27	0.02	0.18
Large	Large	Substantial	-0.20	0.25	-0.29	0.12
Large	Large	Slight	-0.05	0.77	-0.27	0.09
Large	Large	Zero	0.17	0.76	-0.23	0.27
Large	Medium	Substantial	-0.13	0.76	-0.02	0.20
Large	Medium	Slight	-0.01	1.16	-0.03	0.14
Large	Medium	Zero	-0.02	1.02	0.01	0.14
Large	Small	Substantial	-0.20	1.08	0.21	0.17
Large	Small	Slight	-0.05	1.61	0.30	0.25
Large	Small	Zero	-0.02	1.50	0.18	0.15
Medium	Truncated	Substantial	-0.11	1.18	0.19	0.34
Medium	Truncated	Slight	0.06	1.57	0.28	0.42
Medium	Truncated	Zero	0.00	1.41	0.16	0.28
Medium	Large	Substantial	-0.08	0.45	-0.25	0.24
Medium	Large	Slight	-0.03	0.74	-0.33	0.15
Medium	Large	Zero	0.00	0.80	-0.21	0.18
Medium	Medium	Substantial	-0.18	0.84	0.00	0.22
Medium	Medium	Slight	-0.07	1.26	0.06	0.20
Medium	Medium	Zero	0.07	1.40	0.19	0.28
Medium	Small	Substantial	-0.18	1.34	0.38	0.37
Medium	Small	Slight	-0.03	1.73	0.42	0.43
Medium	Small	Zero	-0.05	1.66	0.32	0.29
Small	Truncated	Substantial	-0.01	1.41	0.27	0.34
Small	Truncated	Slight	0.01	1.49	0.30	0.32
Small	Truncated	Zero	0.00	1.63	0.32	0.36
Small	Large	Substantial	-0.03	1.00	-0.12	0.19
Small	Large	Slight	0.04	0.92	-0.07	0.20
Small	Large	Zero	0.03	0.94	-0.18	0.16
Small	Medium	Substantial	-0.03	1.37	0.12	0.31
Small	Medium	Slight	-0.01	1.61	0.17	0.34
Small	Medium	Zero	-0.05	1.43	0.24	0.27
Small	Small	Substantial	0.04	1.51	0.40	0.40
Small	Small	Slight	0.00	1.59	0.39	0.39
Small	Small	Zero	0.06	1.90	0.44	0.44

Note. The corresponding table for Scenario B is presented in Appendix E.

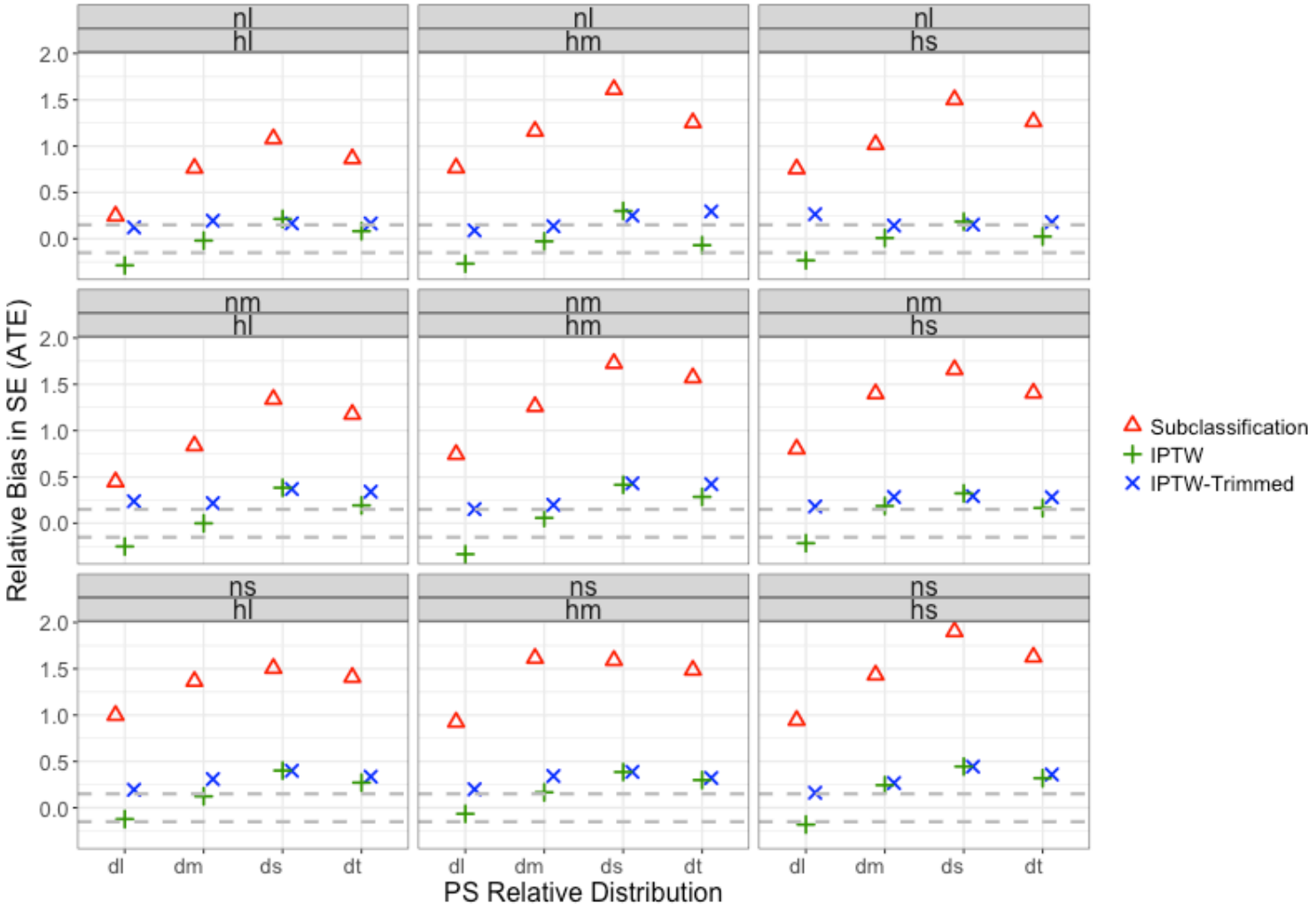


Figure 15. Relative bias in SE for ATE methods (Scenario A). The plot for scenario B is presented in Appendix D. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects. The grey dotted lines indicate the thresholds between *use with caution* and *untrustworthy* (±15%).

4.3 PS weighting

The results for PS weighting, including balance, accuracy, precision, and relative bias in SE estimate are interpreted in this section for both WBO and IPTW. The interpretation focused on PS weighting *without* trimming. Section 4.3.3 briefly presents the results for PS weighting with trimming and explains why it is not a good idea.

4.3.1 Balance

After PS weighting, the balance of the covariates between the two groups was largely improved. Similar to PS matching, which has a great reputation in obtaining desirable balance, the SMD values were always below 0.25 and were mostly below the strict criterion of 0.1. The only situation when the imbalance was slightly large (still to an acceptable degree) was with large difference in PS relative distributions (see Figure 5).

4.3.2 Accuracy of ATT and ATE estimation (without trimming)

As is shown in Figure 6 and Table 5, the ATT estimated via WBO was robust to PS relative distributions, heterogeneous treatment effects, and relative sample sizes between the two groups. WBO constantly produced accurate ATT estimates regardless of the data conditions simulated in the study. This can be explained by the final weighted PS distributions used to calculate the TE estimate generated by weighting. As was illustrated in Chapter 2, with PS weighting, each subject is weighted such that the weighted PS distribution either reflects the true PS distribution of the treatment group with WBO, or the PS distribution of the overall population (including both groups) with IPTW. Figure 16 shows what the weighted PS distributions look like as compared to the original PS distribution with a large difference between the two groups (i.e., small overlap).

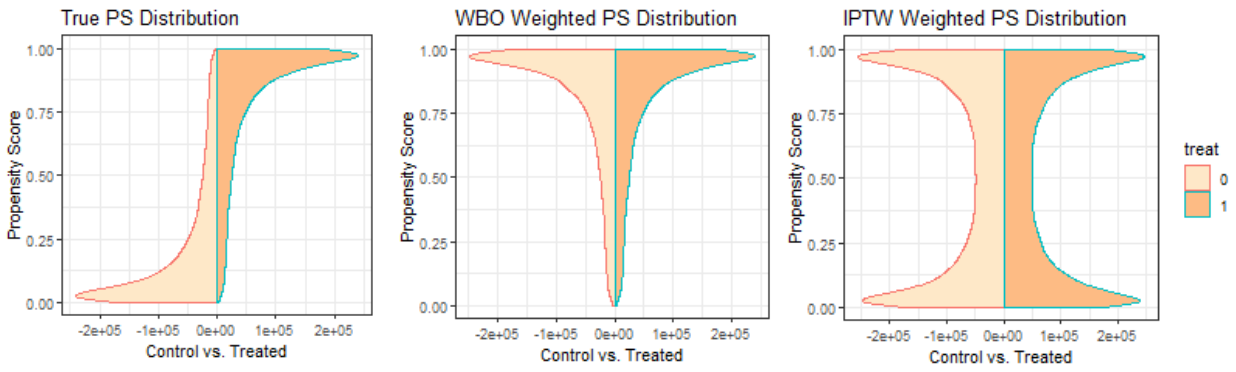


Figure 16. Original vs. weighted PS distributions

After weighting, the PS distributions of the treatment and control groups were almost identical, with the WBO weighted distributions perfectly mimicking the true PS distribution of the treatment group and the IPTW weighted distributions perfectly mimicking the true PS distribution of the full population. This is true regardless of what the original PS distributions look like and it explains why PS weighting is robust to PS relative distributions.

The fact that the weighted PS distribution perfectly reflect the true PS distribution of the population also explains why PS weighting is robust to heterogeneous treatment effects. Because different propensity scores are associated with different treatment effects, a PS method is usually sensitive to TE heterogeneity in that the estimated PS distribution is different from the population. A good example is PS 1:1 matching. By removing cases, the weights taken by subjects that represent different treatment effects are changed and so will be the overall TE estimates. With PS weighting, the weighted PS distribution is almost identical to the true PS distribution, regardless of differences in treatment effects across propensity scores.

4.3.3 Accuracy of ATT and ATE estimation (with trimming)

Unlike PS weighting without trimming, PS weighting with trimming is sensitive to PS relative distributions. With PS trimming, the WBO weights that were greater than 10 or smaller than 0.1 were set to be 10 and 0.1 in the control group, while those in the treatment group

remained the same. In the simulated data, take the condition of large PS relative distribution as an example, about 1.5% of high weights and 42.9% of low weights were trimmed in the control group. Trimming low weights seemingly affected more subjects in the data and might have been much more influential on the TE estimates, however, the actual numbers of subjects trimmed off due to high weights and added in by bringing up the low weights respectively represented 24.5% and 2.7% subjects in the treatment population, indicating that the trimmed high weights that corresponded with higher propensity scores were the dominate reason for the bias in the ATT estimate (Figure 16). To illustrate with more detail, in the current simulation, higher propensity scores were associated with lower ITEs (i.e., $r = -.65$, $p < .001$), trimming more subjects with higher propensity scores brought up the weighted mean of the control group. Since TE estimates were calculated as the weighted mean difference in the outcome between the treatment and control groups, increasing the weighted mean of the control group resulted in a negative bias and thus positive relative bias given that the true treatment effect was negative (Figure 6). How trimming inflated the bias in the ATE estimates with IPTW follows the same logic and thus is not repeated.

The direction of the relative bias is data-dependent. For example, the relative bias might be negative if more weights on the lower end were trimmed; the relative bias would be small only if the higher and lower weights were equally trimmed, which is almost always impossible in the real world. Since PS trimming changes the weighted PS distributions that are almost perfectly representative of the population, it is very likely to cause bias in the TE estimates, especially with greater difference in PS relative distributions where more weights will be trimmed due to more extreme propensity scores. PS trimming is thus not recommended.

4.3.4 Precision

The standard errors for PS weighting were most sensitive to PS relative distributions compared to the other methods (Table 8, Table 11, Figure 8, and Figure 14). When the difference in relative PS distribution was large, the precision generated by PS weighting (both WBO and IPTW) were drastically worse than the other methods, indicating that the TE estimates were not generalizable to other samples in the population even though PS weighting obtained desirable accuracy. Although trimming is not recommended due to unacceptable bias, unsurprisingly, it showed the best precision for most conditions compared to the other PS conditioning methods. After tweaking ways of setting cutoffs, PS weighting with trimming may yield some balance between balance and precision, however, this is beyond the scope of this study. Again, while the precision of the TE estimates from the PS weighting methods indicates whether the results are generalizable, it is not used as a key performance metric to decide whether or not to choose a PS conditioning method with a particular dataset, as it may not always be correctly estimated in the first place. Instead, the relative bias in SE estimate does the work.

4.3.5 Relative bias in SE estimate

Both WBO and IPTW had lower relative biases in SE estimates relative to other PS conditioning methods (i.e., WBO compared to matching and subclassification for ATT, IPTW compared to subclassification for ATE; Table 9, Table 12, Figure 9, and Figure 15). There were more data conditions for PS weighting when the relative bias in SE estimate was acceptable (i.e., either trustworthy or use with caution).

No clear pattern was observed in terms of which simulated conditions were associated with acceptable relative bias in SE. This is understandable because unlike the other performance

metrics, this metric is depending on two variables – the empirical SE of the model and the average SE estimate.

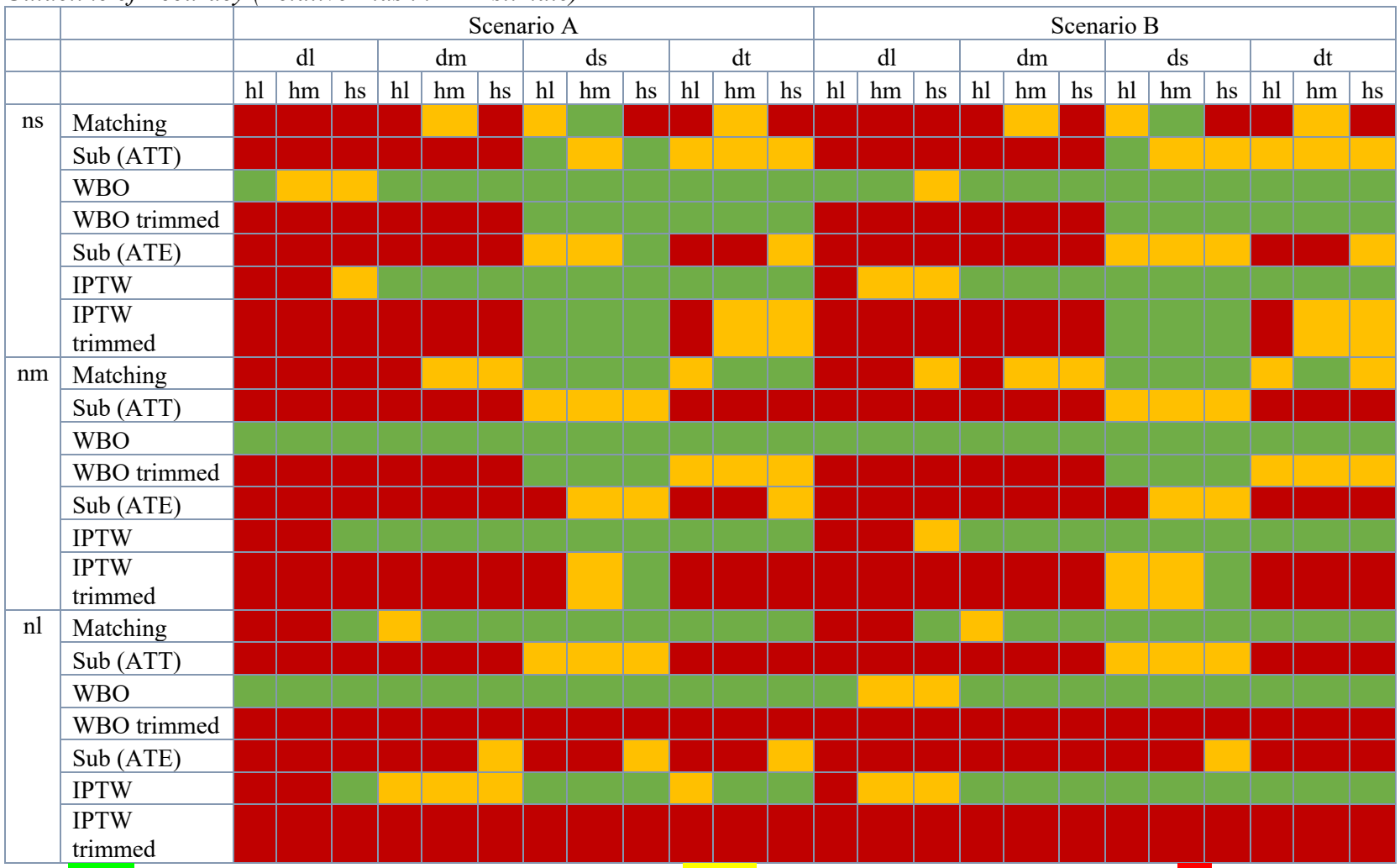
Similar to precision, PS weighting with trimming provided low and stable relative bias in SE estimates. However, because the TE estimates were biased in the first place (interpreted in Section 4.3.3), the results here were not of interest of the study.

4.4 Guideline Tables

In summarizing the performance metrics, relative bias in TE and SE estimates in particular, guideline tables were generated to show which PS conditioning methods were trustworthy or untrustworthy in which data condition (Table 13 and Table 14). A hybrid guideline table (Table 15) was further put together combining the results presented in Table 13 and Table 14, providing a one-stop guideline for which PS method to choose. Practitioners can easily refer to these tables to decide which PS conditioning method produces the most trustworthy TE and SE results under the data condition of interest. The thresholds for each of the three categories (i.e, “good to use”, “use with caution”, and “untrustworthy) were discussed in Section 3.5 and presented in Table 4. The following chapter (Chapter 5) illustrated, via an empirical analysis, how to use the guideline tables effectively.

Table 13

Guideline of Accuracy (Relative Bias in TE Estimate)



Note. Green = good to use (relative bias in TE estimate < 5%), yellow = use with caution (>= 5% & < 10%), red = untrustworthy (>= 10%). nl(nm/ns) = large(medium/small) sample size in the control group. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions; hl(hm/hs) = substantial(slight/zero) heterogeneous treatment effects.

Table 14

Guideline of Precision (Relative Bias in SE Estimate)

		Scenario A												Scenario B											
		dl			dm			ds			dt			dl			dm			ds			dt		
		hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs
ns	Matching																								
	Sub (ATT)																								
	WBO																								
	WBO trimmed																								
	Sub (ATE)																								
	IPTW																								
	IPTW trimmed																								
nm	Matching																								
	Sub (ATT)																								
	WBO																								
	WBO trimmed																								
	Sub (ATE)																								
	IPTW																								
	IPTW trimmed																								
nl	Matching																								
	Sub (ATT)																								
	WBO																								
	WBO trimmed																								
	Sub (ATE)																								
	IPTW																								
	IPTW trimmed																								

Note. Green = good to use (relative bias in SE estimate < 10%), yellow = use with caution (>= 10% & < 15%), red = untrustworthy (>= 15%). nl(nm/ns) = large(medium/small) sample size in the control group. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions; hl(hm/hs) = substantial(slight/zero) heterogeneous treatment effects.

Table 15

Overall Guideline (Combination of Accuracy and Precision)

		Scenario A												Scenario B											
		dl			dm			ds			dt			dl			dm			ds			dt		
		hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs	hl	hm	hs
ns	Matching					t		t	t			t						t		t	t			t	
	Sub (ATT)							t	t	t	t	t	t							t	t	t	t	t	t
	WBO	t	t	t			t	t	t	t	t		t	t	t	t				t	t	t		t	t
	WBO trimmed	s	s	s					t	t	t		t	s	s	s	s	s	s	t	t	t	t	t	t
	Sub (ATE)							t	t	t			t							t	t	t			t
	IPTW	s	s	t	t		t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
	IPTW trimmed							t	t	t		t	t							t	t	t		t	t
nm	Matching		s			t	t		t	t	t	t	t			t		t	t	t	t	t	t	t	t
	Sub (ATT)							t	t	t										t	t	t			
	WBO				t		t	t	t	t	t	t	t		t	t	t	t		t	t	t	t	t	t
	WBO trimmed			s	s	s			t	t	t	t	t	s	s	s				t	t	t	s	t	t
	Sub (ATE)								t	t			t								t	t			
	IPTW			t			t	t	t	t	t	t	t			t			t	t	t	t	t	t	t
	IPTW trimmed								t	t						s				t	t	t			
nl	Matching			s	s	t	t		t	t	t	t	t			t	t	t	t	t	t	t	t	t	t
	Sub (ATT)							t	t	t										t	t	t			
	WBO				t	t	t	t	t	t	t	t	t		t	s	t	t	t		t	t	t	t	t
	WBO trimmed	s	s					s			s			s		s				s			s		
	Sub (ATE)						t			t			t									t			
	IPTW			t	s	s	s	t	t	t	s				t	t				t	t	t			
	IPTW trimmed	s	s			s	s									s						s	s		

Note. Each color represents for the combination results of TE and SE estimates. Green = both are good to use, yellow = both use with caution, red = both untrustworthy. The other colors indicate inconsistent relative bias in TE and SE estimates, with **t** = TE estimate falls in a better category, **s** = SE estimate is better. Olive = good to use & use with caution, orange = use with caution & untrustworthy, purple = good to use & untrustworthy. nl(nm/ns) = large(medium/small) sample size in the control group. dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions; hl(hm/hs) = substantial(slight/zero) heterogeneous treatment effects.

Chapter 5. Empirical Illustration

In addition to the methodological study presented in Chapters 3 and 4, I conducted an empirical analysis to demonstrate the PS procedure including an extra step of checking PS relative distributions and choosing the most appropriate PS conditioning method. The analysis was conducted using the data selected from the Early Childhood Longitudinal Study Kindergarten Class of 1998-99 (ECLS-K) released by the National Center for Education Statistics (NCES, Tourangeau et al., 2015), assessing the effect of having home computers (COMPUTER) on first grade students' math scores. To be consistent with the simulation study where the data were generated following a single-level structure, the initial data analysis was conducted as if the ECLS-K:2011 data were collected following a simple random sampling strategy. After the procedure demonstration, in order to provide education decision makers an unbiased and precise inference, a second analysis was conducted accounting for the complex sampling structure of the data. This chapter includes a data description, how the two data analyses were implemented, and the corresponding results.

5.1 Data

Data were selected from ECLS-K:2011, released by NCES, which was designed to provide comprehensive and reliable data that contain repeated observations of a nationally representative sample of students, their families, teachers, and schools across the United States. The ECLS-K:2011 data employed a three-stage sampling design. Specifically, geographic areas (i.e., counties or county groups) were first selected as primary sampling units (PSU); within each of the selected PSUs, public and private schools with kindergarten programs were selected, and children were then sampled from the selected schools. In the first two stages, both geographic areas and schools were selected with probability proportional to measures size. Sampling

weights were generated to compensate for differential probabilities of selection and to adjust for the potential effects due to nonresponse (Tourangeau et al., 2015).

This analysis was limited to children who had a non-zero base weight *W4C4P_40* (their parents completed surveys conducted during the fall [2010, wave 1] and spring [2011, wave 2] of children's kindergarten year, the spring [2012, wave 4] of their year in the first grade), complete data on the treatment variable *P2HOMECM* (having access to a computer at home in the spring semester of the kindergarten year), and complete data on the outcome variable *X4MSCALK1* (math scores in the spring semester of first grade). This analytic sample includes 9,209 students from 1,398 schools, among which only 75.5% students had computer(s) at home.

Among the PS conditioning methods, nearest neighbor matching for the estimation of ATT assumes that the number of treated cases is no larger than the controlled cases. This is because it usually matches controlled cases to treatment cases and discards the controls that are not matched to the treatment group (Stuart, 2010); if the sample size in the control group is smaller than the treatment group and therefore treated cases are discarded to yield the balance between the two groups, it will cause "bias due to incomplete matching" (Rosenbaum & Rubin, 1985). Considering that PS matching was one of the major conditioning methods and had the potential to be selected as the best approach in the analyses, the treatment in this data analysis was switched from "having a computer at home" to "not having computer at home". With the new treatment definition, the sample size for the control group is about 3 times of that for the treatment group.

5.2 Methods for Empirical Illustration

The empirical illustration followed the new six-step PS procedure suggested in the simulation study, including a new step of checking the PS relative distribution and selecting the

most appropriate PS conditioning method for the corresponding PS distributions. This section describes how each of the six steps (i.e., variable selection, PS estimation, PS relative distribution checking and conditioning method selection, PS conditioning, balance checking, and TE estimation) were implemented in detail.

5.2.1 Variable selection and missing data

In this empirical illustration, the effect of COMPUTER was examined following the typical five-step PS procedure, with the application of the PS conditioning methods studied in the simulation. A challenge that did not exist in the simulation study was to select appropriate pretreatment covariates for the PS model, as the true model was not available. Selection of pretreatment covariates for the PS model was based on literature regarding educational technology, data structure, and statistical rules suggest by previous studies.

Literature on educational technology shows that income and other socioeconomic factors (e.g., education and occupation) were identified as strong predictors of the adoption of home computers (Becker, 2000; Hoffman & Novak, 1998). For example, according to Becker (2000), at the end of 1998, 91% of children living with a parent having at least a Master's degree had home access to a computer, compared to only 16% of children with parents who did not graduate high school. Similarly, the disparities of COMPUTER between families whose household incomes were under $20,000 per year and those higher than $75,000 per year are also substantial. Other than socioeconomic factors, demographic categories such as ethnicity, parents' marital status, and number of children in household can also explain the disparities on COMPUTER (Becker, 2000; Carroll, Rivara, Ebel, Zimmerman, & Christakis, 2005; Hoffman & Novak, 1998).

Selection into the treatment COMPUTER can be relatively well explained by the individual-level variables described above. However, in this data setting where students are nested within different schools, it is not sufficient to solely consider the individual-level descriptive structural characteristics of students and their families. The school environment, such as school type and school resources, could also contribute to the treatment and thus were included.

In addition to the conceptual considerations, covariates were selected also following the statistical suggestions summarized in Stuart and Rubin (2007), that is, selecting covariates that are related to the treatment assignment, including a large set of covariates for greater power, and not selecting covariates that are affected by treatment assignment.

As such, a total number of 28 covariates from the year of kindergarten were selected for the PS model, aiming to incorporate as much potential confounding as possible. The number of covariates selected for the research data is substantially greater than the number of covariates generated in the simulation study. This is because the simmulated data were intentionally simplified to focus on the research interest related to the relative PS distributions. The number of covariates should not affect the comparability between the empirical illustration and the simulation results, as long as the empirical distribution is evaluated appropriately. The descriptions and coding schemes of each covariate are presented in Table 16.

Table 16

Selected (ECLS-K:2011) Variables for the Data Analysis

Variable (recoded)	Type	Original Variables in the Data and Description
WEIGHT		Child base weight adjusted for nonresponse associated with child assessment data from both kindergarten rounds and spring first grade, parent data from fall kindergarten or spring kindergarten, and parent data from spring first grade (W4C4P_40).
STRATA		Stratum identifier (W4PF_4STR).
CLUSTER		PSU (cluster) identifier (W4PF_4PSU).
MATH	C	Outcome variable, math IRT scale score (X4MSCALK1).
NOCOMPUTER	D	Treatment variable, students have NO computer at home (P2HOMECM!=1)
EXPECT	D	Parents expect children to at least get a college degree (P1EXPECT=4, 5, 6, 7).
PREMATH	C	Treatment effect predictor, math IRT scale score in the kindergarten year (X2MSCALK1).
SES	C	X12SESL (RENAME TO SES).
PARENTED	D	Parent has at least a college degree (X12PAR1ED_I=6, 7, 8 or X12PAR2ED_I=6, 7, 8).
PARENTOCC	D	Parent's occupation related to science and technology (X1PAR1OCC_I=2,3,5,7,10,11,18; X1PAR2OCC_I=2,3,5,7,10,11,18). Combine the two variables into one dummy variable indicating parent's occupation related to computer technology (e.g., scientists, engineers, and technologists).
INCOME	C	Household income intervals (X2INCCAT_I, assuming a continuous scale)
POVERTY	D	X2POVTY=1
MARRIAGE	D	Bio-parents are married (P2BIOMRY=1).
Race	D	White (X_RACETH_R=1).
	D	African American (X_RACETH_R=2).
	D	Hispanic (X_RACETH_R=3, 4).
	D	Asian (X_RACETH_R=5).
TVRULE	D	Rule on hours of watching TV each week (P2TVRUL3=1).
VISION	D	Students have difficulty seeing objects in the distance or letters on paper (P2SIGHT=1). Use the kindergarten variable to reduce the effect of treatment on the covariate.
BEDRULE	D	Children go to bed the same time (P2GOTOBD=1).
HTOTAL	C	Number of children (aged<18) in each household (X2LESS18).
LIBRARY	D	Visited library in the past month (P2LIBRAR=1).
ATHLET	D	Students participate in athletic activities outside of school hours (P2ATHLET=1).
NONENGLISH	D	Students speak non-English language at home (C2ENGHM=1).

(Continued)

Table 16 (Continued)

Variable (recoded)	Type	Original Variables in the Data and Description
LAB	D	Whether computer lab meets needs (S2COMPOK=5).
PRIVATE	D	Private school (X2PUBPRI=2).
URBAN	D	Urban school (X2LOCALE=1).
SUBURBAN	D	Suburban school (X2LOCALE=2).
FREELCH	D	50% of students eligible for Free lunch (P2LUNCHS).
SCHSIZES	D	Small school size (X2KENRLS=1).
SCHSIZEL	D	Large school size (X2KENRLS=4, 5).
MEANSES	C	Mean SES (aggregate X12SESL).
SDSES	C	The standard deviation of SES scores within each school.

Note. C=continuous variable; D=dichotomous variable. All PS pretreatment covariates were selected from fall [2010, wave 1] or spring [2011, wave 2] of children's kindergarten year to ensure that the covariates were not affected by the treatment. The first 3 variables in the list WEIGHT, STRATA, and CLUSTER were selected to accommodate the complex sampling structure.

The ECLS-K:2011 data were not without missingness. For most of the dummy variables, missingness were treated as 0 without losing useful information. Take the parent occupation variables as an example, the occupation question was not applicable for 0.2% of the samples who did not have any job at that time; 3% of the parents chose "not ascertained" because they were unemployed, retired, or the occupation was unclassifiable. In this case, it is reasonable to classify the missing subjects as not having an occupation related to science and technology. Another example is whether parents expect their child to at least get a college degree. About 30 parents either refused to answer the question or indicated that they did not know. Because they did not clearly state that they had an expectation for their children to get at least a college degree, they were coded as 0. Other missing imputation techniques included mean imputation for continuous variables (i.e., for 22 math scores that were missing in the previous semester, 19 schools in which SES was not available) and mode imputation for count/categorical variables (i.e., the number of children in the family for 33 students was "not ascertained" and the mode of 2 for this variable was imputed). The limitation of data imputation with central tendency is that it tends to reduce the variance of the imputed variables and therefore may lead to inflated Type II error rate. Given that only a small number of observations had to be imputed, mean/mode imputation likely had little effect on the resulting variance estimates.

5.2.2 PS estimation

Based on the 28 covariates selected from the first step, I constructed a conceptual diagram as is shown in Figure 17. Following the conceptual framework, a logistic

regression (shown in Equation 22) was fit to estimate the propensity scores for each observation:

$$\text{logit}(\hat{e} \mid T = 1) = \hat{\beta}_0 + \hat{\beta}_1 PREMATH + \hat{\beta}_2 EXPECT + \hat{\beta}_3 SES + \ldots + \hat{\beta}_{28} SDSES\,, \quad (22)$$

where $\hat{e}$ represents the PS estimates; $\hat{\beta}_0 - \hat{\beta}_{28}$ are the estimated coefficients for each of the pretreatment covariates presented inTable 16. With the PS estimates, the PS relative distribution was checked to decide which PS conditioning method would produce the best performance in terms of accuracy and relative bias in SE for TE estimation.

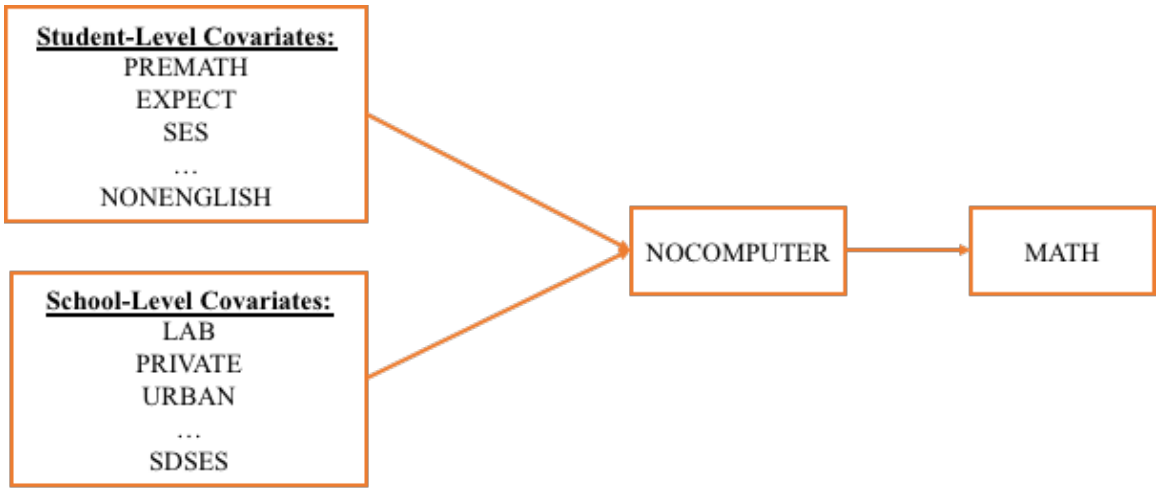


Figure 17. The conceptual model for the empirical study. A complete list of covariates is shown in Table 16.

5.2.3 PS distribution checking and conditioning method selection

Based on the PS model, PS estimates as well as the descriptive statistics of the relative PS distributions of the real data were calculated. The statistics related to the PS relative distributions (i.e., overlap, mean, variance, skewness, and kurtosis) are presented in Table 17, and the distributions are visualized in Figure 18. Although the statistics for the real data do not perfectly match any of the simulated conditions shown in Table 2, the distributions look very similar to the medium relative distributions. Having the sample

size in the control group (i.e., having a computer at home) about 3 times of the sample size in the treatment group (i.e., not having a computer at home), it falls in the medium sample size category in the simulated conditions. With respect to heterogeneity in treatment effects, medium heterogeneity was assumed to reflect reality because homogeneity of treatment effect tends to be a strong assumption to make with real data (Wyss, Glynn & Brookhart, 2014; Xie, Brand, & Jann, 2012). In addition to the relative PS distribution, sample size, and heterogeneity, the PS model was assumed to be correctly specified (i.e., consistent with Scenario A) given that the simulation results for Scenarios A and B were similar. By checking the guideline tables (Table 13 and Table 14), WBO and IPTW produced the best accuracy in TE and SE estimates for the condition of *medium PS relative distribution–medium sample size in the control group–medium heterogeneity in the treatment effects–Scenario A*, and thus, were selected as the conditioning methods for this analysis, producing ATT and ATE estimates respectively.

Table 17

Descriptives of the Relative PS Distributions of the Empirical Data

Overlap	Mean		Variance		Skewness		Kurtosis	
	Treat	Control	Treat	Control	Treat	Control	Treat	Control
0.46	0.35	0.21	0.03	0.02	0.10	1.19	1.97	3.80

Note. The overlap quantities are calculated as the proportion of the intersection of the two PS distributions over the union of the two distributions.

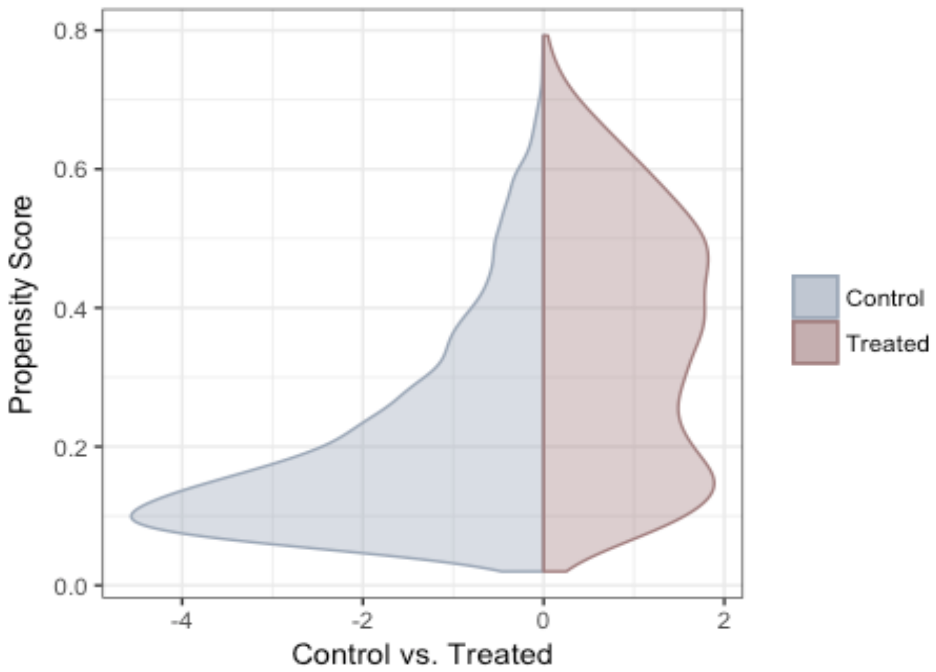


Figure 18. The relative PS distributions of the treated and control groups for the empirical data.

5.2.4 PS conditioning

The WBO and IPTW weights were calculated based on Equations 6 and 7. For the reasons indicated in Section 4.3.3 regarding how weights were trimmed and how that introduced more bias in the TE estimates, weight trimming was not considered in the data analysis. Before using the weights for TE estimation, a diagnostic step was implemented to check the balance, that is, whether the conditioned data were comparable between the treatment and control groups.

5.2.5 Balance checking

For the conditioned data, balance was checked via the performance metric SMD defined in Equation 8. The balance presented in Figure 19 indicated that both WBO and

IPTW did a good job yielding comparable control and treated objects, on aggregate. The analysis was then proceeded for TE estimation.

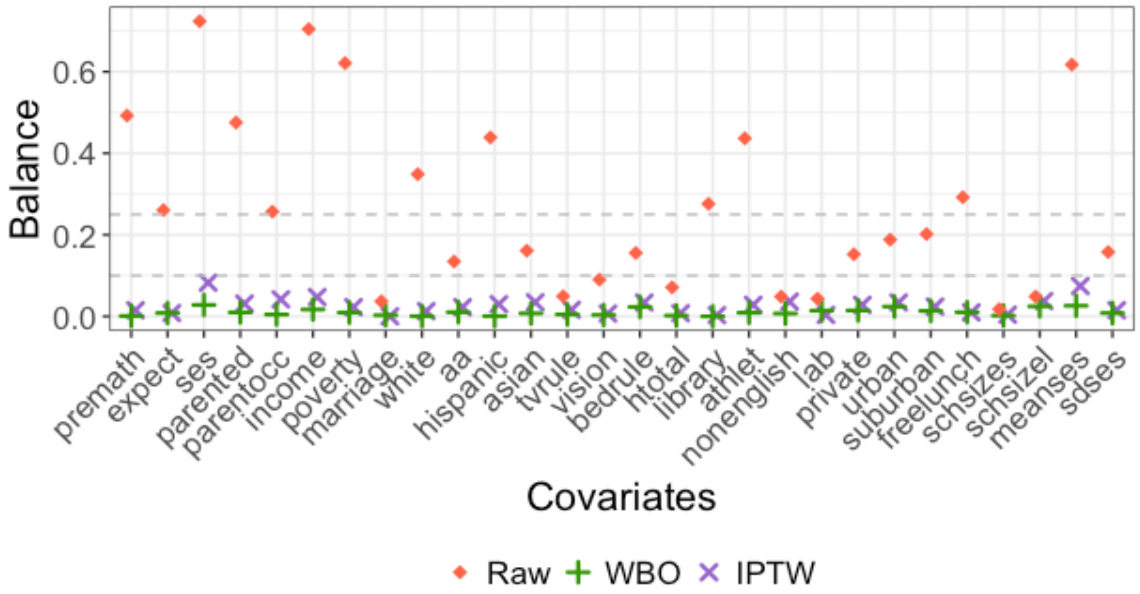


Figure 19. Standardized mean difference (SMD) for all PS conditioning models for the empirical data.

5.2.6 TE estimation

According to the conceptual diagram shown in Figure 17, the TE model can be statistically expressed as:

$$\hat{Y} = \hat{\gamma}_0 + \hat{\gamma}_1 T\,, \tag{23}$$

where $\hat{\gamma}_0$ is the intercept; $\hat{\gamma}_1$ represents for the TE estimate and is the focus of the research. This was a weighted model using WBO and IPTW weights respectively for the estimation of ATT and ATE. The results for a naïve model without PS adjustment for confounding was also implemented for comparison.

The results presented in Table 18 indicate that for the 1st grade students in the US who did not have computers at home, their scores would have increased by only 0.5 points (out of a full score of 100) if they did have computers at home; for all 1st grade students in the US, not having computers at home would make their math scores drop by 0.004 points compared to if they did have computers at home. The two interpretations respectively represented the results for ATT and ATE. Neither of the results, however, was statistically significant.

Table 18

The TE Estimates and Standard Errors for the First Empirical Analysis

		ATT	ATE
	Naïve	WBO	IPTW
TE	-6.63***	-0.50	-0.004
SE	0.31	0.38	0.39

Note. p < .05, ** p < .01, *** p < .001. To find the p-values for the PS weighting methods, the squared z-statistics (TE estimates divided by the standard errors calculated via the robust sandwich variance estimator) were calculated and compared to a chi-squared distribution on one degree of freedom.

5.3 A Second Data Analysis

While the simulation design only focused on data with a simple random sampling structure, as described in Section 5.1, the large-scale education data ECLS-K:2011 were collected following a three-stage complex sampling design (Tourangeau et al., 2015) that involves clustering, stratification, and disproportional selection. To simply demonstrate a complete process of the six-step PS procedure recommended in the simulation study, the first analysis was conducted as if the ECLS-K:2011 data were selected following a simple random sampling strategy. However, failure to address any of the complex sampling features could directly bias either the TE estimate or its standard error. For example, standard errors can be deflated if not considering clustering whereas inflated if

stratification is ignored; deflation in standard errors as well as biased parameter estimates will occur if the probability of unequal selection is not taken into account (Heeringa, West, & Berglund, 2010; Pfeffermann, 1993). The purpose of this second analysis, therefore, was to make adjustments to the TE model to account for the complex sampling structure of the data, in order to obtain a more accurate and generalizable TE estimate. This better inference would be valuable for education policy makers who are interested in making decisions based on how not having a home computer would affect 1st grade students' math achievement.

As a supplement to the first analysis, the second data analysis directly adopted the conditioned samples via the selected PS conditioning methods (i.e., WBO and IPTW) in the first analysis based on the single-level PS model. In fact, using a single-level PS model in a complex sampling data setting is a reasonable idea, because the PS estimates are only used to obtain balance between the treatment and control groups rather than to directly make inferences about the PS model in the population (Zanutto et al., 2005). According to the simulation study by An and Stapleton (2016), the balance yielded via a single-level PS model was very comparable to multiple other methods that accounted for the complex data structure. That being said, the only difference between the first and the second analyses was how the treatment effects were estimated.

Researchers have explored two types of methods in the application of PS methods for TE estimation: model-based techniques to consider the clustering effect via multilevel models (Hong & Raudenbush, 2006; Hong & Yu, 2007, 2008; Kelcey, 2011; Thoemmes & West, 2011) and design-based techniques to account for the features of complex sample data (Dugoff et al., 2014; Hahs-Vaughn, 2015; Zanutto, 2006; Zanutto et al.,

2005). For this analysis, the TE model was selected following the recommendations by An and Stapleton (2016), that is, fitting a design-based model for the TE estimation using the sampling weight * IPTW or WBO weights as the final weight for the TE model.

The results for these two models are presented in Table 19. To interpret the results in more details, for the 1st grade students who had no home computer, their scores would have been .62 points higher if they did have computers at home. For all 1st grade students in the US, having computer at home would have increase their math scores by .09. However, neither of the results were statistically significant. Therefore, there is no sufficient evidence to support that not having a computer at home would affect 1st grade students' math achievement in the US. This inference can be used to support decision making. For example, for policy makers who are considering distributing free computers to 1st grade students in the US who do not have any computers at home to improve their math scores, they may reconsider the idea because the investment would not be statistically effective in improving their math scores.

Table 19

The TE Estimates and Standard Errors for the Second Empirical Analysis

	WBO	IPTW
TE	-.62	-.09
SE	.15	.83

Note. * $p < .05$.

5.4 Summary

In Chapter 5, I introduced the designs and results for two empirical analyses. The first one was a complete illustration of how to check PS relative distribution and select the best PS conditioning method in practice. The second data analysis further accounted for the complex sampling structure in the TE model and provided to the education policy

makers unbiased causal inference about how not having a home computer affects 1st grade students' match achievement. The results showed that there was no sufficient evidence that not having computers would have affected 1st grade students' math achievement.

Chapter 6. Discussion & Recommendation

The study extended the propensity score literature by investigating whether and how the relative PS distributions affect the performance of PS conditioning methods (i.e., matching, subclassification, and PS weighting) for the purpose of generating appropriate causal estimates. In addition to PS relative distributions, two other primary factors were manipulated in the simulation study, including heterogeneity in the treatment effect and sample size in the control group. A final manipulation of PS specification in two scenarios did not produce obvious differences in the performance metrics and thus was not discussed in detail in the study. Following the simulation study, an empirical illustration was completed as a demonstration of how to utilize the results from the simulation study in practice. This chapter covers a summary of key findings from the simulation study, limitations, as well as possible extensions for future research.

6.1 Summary of Key Findings

The goal of the study was to answer the questions that appear to have never been directly discussed in the literature, that is, whether the quality of the TE estimates via different PS conditioning methods are affected by the relative PS distributions between the treated and control groups, and whether the relation between the relative distributions and performance of the PS conditioning methods differs across levels of TE heterogeneity. The key findings of different PS conditioning methods (including PS matching, subclassification for ATT and ATE, as well as IPTW and WBO) are summarized as follows. PS weighting with trimming is not recommened due to the fact that it introduced bias to the TE estimates, and thus is not discussed here.

First, almost all PS conditioning methods yielded acceptable balance across different data conditions, except when the PS relative distribution was large with medium and large sample sizes in the control group. Balance was not sensitive to heterogeneous treatment effects at all due to the fact that the heterogeneous treatment effects reflect the relations between the outcome and the propensity scores obtained via covariates, rather than the characteristics of the covariates within each treatment group, which actually determine what the balance metric should look like. In addition to SMD used in this study, the ratios of the variances for each covariate between the treated and control groups can be another option to measure balance (Austin, 2009).

Second, PS weighting (both IPTW and WBO) almost constantly produced accurate TE estimates, regardless of the data conditions simulated in the study. Matching was sensitive to PS relative distributions, with more TE bias associated with larger differences in PS relative distributions unless there was sufficient sample in the control group such that all treated cases were retained for TE estimation. Therefore, in the simulated conditions, the more cases in the control group, the better the accuracy. For matching, in addition to the sample size in the control group, the level of heterogeneity in treatment effects also contributed to how many cases were removed from the control group, for the reason that the PS estimates tend to be more extreme with a higher level of TE heterogeneity. Therefore, the factor of heterogeneous treatment effects also affected the accuracy of ATT estimates. Similar to PS matching, subclassification (for both ATT and ATE) was also sensitive to PS relative distributions, however, with lower accuracy in TE estimate associated with larger sample size in the control group. This is because the sample size in the control group determined how many treated cases were classificed into

each subclass. Specifically, with a larger controlled sample size, a smaller number of cases (sometimes too small) would be assigned to the treatment group for certain subclasses, for the reason that the classification was made based on the *overall* sample rather than the *treated sample only*. This will be further discussed in Section 6.2. Subclassification for ATT was surprisingly not sensitive to heterogeneous treatment effects, due to the fact that the estimated treatment effects were highly correlated (r = .61) with the generated true treatment effects and thus the bias was almost consistent across different heterogeneity levels. Since this result possibly depended on the way the data were generated, the accuracy in ATT estimates for subclassification in the condition of large heterogeneous treatment effects should be considered with caution.

Third, although the PS weighting methods were the best in terms of producing accurate TE estimates, they showed worse precision (i.e., measured via empirical SE) in general compared to the other PS conditioning methods. PS matching was sensitive to large differences in relative PS distributions with respect to precision and produced the best precision in TE estimates with small sample size in the control group compared to the other methods. Precision for matching however was robust to heterogeneous treatment effects. Subclassification (both ATT and ATE) followed similar pattern as matching in terms of precision. PS weighting (i.e., both WBO and IPTW) showed larger SE in general compared to the other methods. While the empirical SE is informative of how generalizable the results are, it is not recommended as a performance measurement metric, because the SE estimate may not even reflect the true SE of a model (i.e., SE estimates untrustworthy). Therefore, the realtive bias in SE estimates is further discussed

in the following paragraph as another quality measure for different PS conditioning methods.

In addition to considering whether the parameter estimates are generalizable, we care more about whether the estimates are trustworthy (i.e., the accuracy of the SE estimates in terms of relative bias). Although weighting had higher SEs in general, the SE estimates tended to be more accurate among all PS conditioning methods. In practice, choosing a method that produces trustworthy result is more desirable than selecting the one that has better but unreliable precision. There were fewer conditions where PS matching generated more trustworthy SE estimates than PS weighting, even though the fluctuation in the SE relative bias was relatively small across conditions. Subclassification, according to the simulation results, never generated SE estimates that were acceptably trustworthy. Because relative bias in SE did not really follow any easy patterns, one needs to refer to the guideline table to decide which method works best under a certain condition.

Finally, the results revealed that when the interaction and quadratic terms in the true PS model were ignored in the PS estimation step, the patterns of the results almost always followed the situation when the PS model was correctly specified. This indicates that, TE estimation is robust to a certain degree of PS misspecification, specifically, ignoring the interaction and quardratic terms. This finding however may not generalize to missing important covariates in the PS model.

6.2 Limitations and Potential Extensions

There are several limitations of the study and potential extensions. First, among the controlled conditions, the PS distribution (and thus the relative PS distribution) is a

data-dependent factor which may differ substantially depending the selected covariates and the distributional features (i.e., mean, variance, skewness, and kurtosis) of the matched data. The shapes of the four relative distribution conditions were empirically rather than quantitatively defined (and the quantities of the overlap were calculated based on the empirical distributions). Although the four relative PS distribution conditions controlled in the simulation already represented different levels of relative patterns of PS distributions, the current design did not cover all scenarios nor produced exact guidelines on how much the treatment effects would be affected by the relative PS distributions. A major consequence could be that when a practitioner is trying to follow the recommendations of this study, he could not match his data to any of the scenarios generated in this design. A possible extension is to collect more empirical PS relative distributions from real data analyses and add more conditions to the simulation. However, a simulation is never going to cover all practical scenarios. An additional point regarding the empirically simulated distributions is that it may seem tempting to use the overlap quantities to evaluate the relative PS distributions, but it should be noted that, if the two distributions have different moments, the results may not generalizable to the empirical research data.

Second, the fact that the PS model misspecification did not produce different patterns in the simulation results could be due to how the misspecification was generated. For the misspecication conditions, I generated a different set of data by incorporating quadratic and interaction terms in the PS data generation model, simply followed the stereotype set by Setoguchi, et al. (2008), without quantifying how big the misspecification was. It could be that the magnitude of the misspecification was not big

enough to produce any significant differences in the results. A possible way to quantify the actual model misspecification is to use the degree of misspecification (DoM) metric proposed by Lenis, Ackerman, and Stuart (2018).

Third, the comparisons made among different methods across the simulated conditions were not based on statistical tests. Instead, thresholds were set as guidance for practice. The justification for not conducting any statistical tests include: (1) Omnibus tests might not be informative given the large number of conditions maniuplated. While multiple comparisons were available to investigate more details such as which conditions differ from others, a larger number of replications might be needed to control for inflated Type I error rate due to multiple testing. (2) Significance test is sensitive to sample size, that is, statistical tests can always be significant with sufficient sample size. That said, the results from the statistical tests might not be generalizable to other real-life scenarios and can be misleading.

Fourth, this study only focused on one popular option of each conditioning method as a first-step investigation, while other options are available and may improve the current results and make the specific conditioning method more competitive compared to others. Take subclassification as an example, a popular way for splitting observations into subclasses and thus was the focus of the study was to use the quintiles of the *overall* sample (Austin, 2011a; Lunceford & Davidian, 2004; Rosenbaum & Rubin, 1983; Stuart, 2010). In the simulation study, this way of splitting classes resulted in extremely small sample size in the treatment group with large PS relative distributions, especially when the sample size in the control group was medium or large. A direct consequence of the imbalanced treated cases across subclasses was that the balance of

covariates between the two groups was not good, indicating that the two groups were not quite comparable in those subclasses. Another consequence was that in certain subclasses where the treated sample size was small, the TE estimates were unreliable and accordingly contributed to an unreliable final TE estimate. Subclassification for ATT tends to be relatively robust to this issue, because the subclasses with few treated cases also take a smaller weight in the overall TE estimate. For ATE, however, because each subclass is equally weighted equal weight, the bias resulting from imbalanced sample size would definitely be reflected in the final ATE estimate. That said, a way to better use PS subclassification, although it is not the most popular in literature, is to stratify based on the sample size in the treatment group. Because the treatment group usually has a smaller sample size, subclassifying by quintiles of the treatment group would help reduce the chance of having extremely small sample sizes within subclasses, and therefore improve the results (i.e., better accuracy in TE and SE estimates). In addition to subclassification, possible ways to improve matching is to minimize "bias due to incomplete matching" (Rosenbaum & Rubin, 1985), via releasing the caliper for 1:1 matching, using 1:n matching instead, etc. It is also possible to improve the quality of TE estimates from PS weighting by optimizing the trimming thresholds. These illuminate directions for future research. Details are not discussed here as they are not the interest of the study.

Another limitation is that due to its two-stage nature (i.e., PS estimation and TE estimation), the propensity score method is subject to the two-stage estimation problem. To illustrate, TE estimation utilized the matched samples constructed based on the PS estimates from the first estimation stage, and therefore the covariance matrix of the

second stage estimator includes noise induced by the first-stage estimates (Karaca–Mandic & Train, 2003). This study, as what most people would do, only considered the two stages independently. However, ignoring the two-stage estimation problem may lead to underestimation in the SE estimates. Future studies are recommended to consider the dependency between the two stages and try to correct the SEs for better inferences.

In addition to the limitations regarding the simulation, the second empirical data analysis in this paper aimed to provide actionable insights to the educators and policy makers. The complex sampling design was considered to reach this goal although the simulation was not conducted under the complex sampling data structure, assuming that the pattern of the simulation results would hold under a complex sampling structure. Comparing the PS conditioning methods with different PS relative distributions in a complex sampling data environment is a recommended area for future research.

6.3 Conclusion

The goal of the study was to find out whether and how different PS relative distributions affect the performance of multiple PS conditioning methods, and how the possible effects would interact with levels of heterogeneity in treatment effects, sample sizes, and PS model specifications. The conclusion is that PS matching and subclassification are both very likely to be sensitive to these conditions depending on what the data look like. PS weighting (without trimming) tends to be robust to a variety of data conditions and produces more accurate and trustworthy TE and SE estimates. An important recommendation for the educational practitioners is that before using the PS methods to make causal inference, try to take an extra step of checking the PS relative distributions and selecting the PS conditioning method that yields the most accurate TE

and SE estimates (see the flowchart in Figure 20). Knowing the relative PS distributions, together with the sample size ratio between the control and treatment groups, as well as an assumption on the level of heterogeneous treatment effects, the practitioners could use the guideline tables created in this study to decide on which PS conditioning method is the most appropriate.

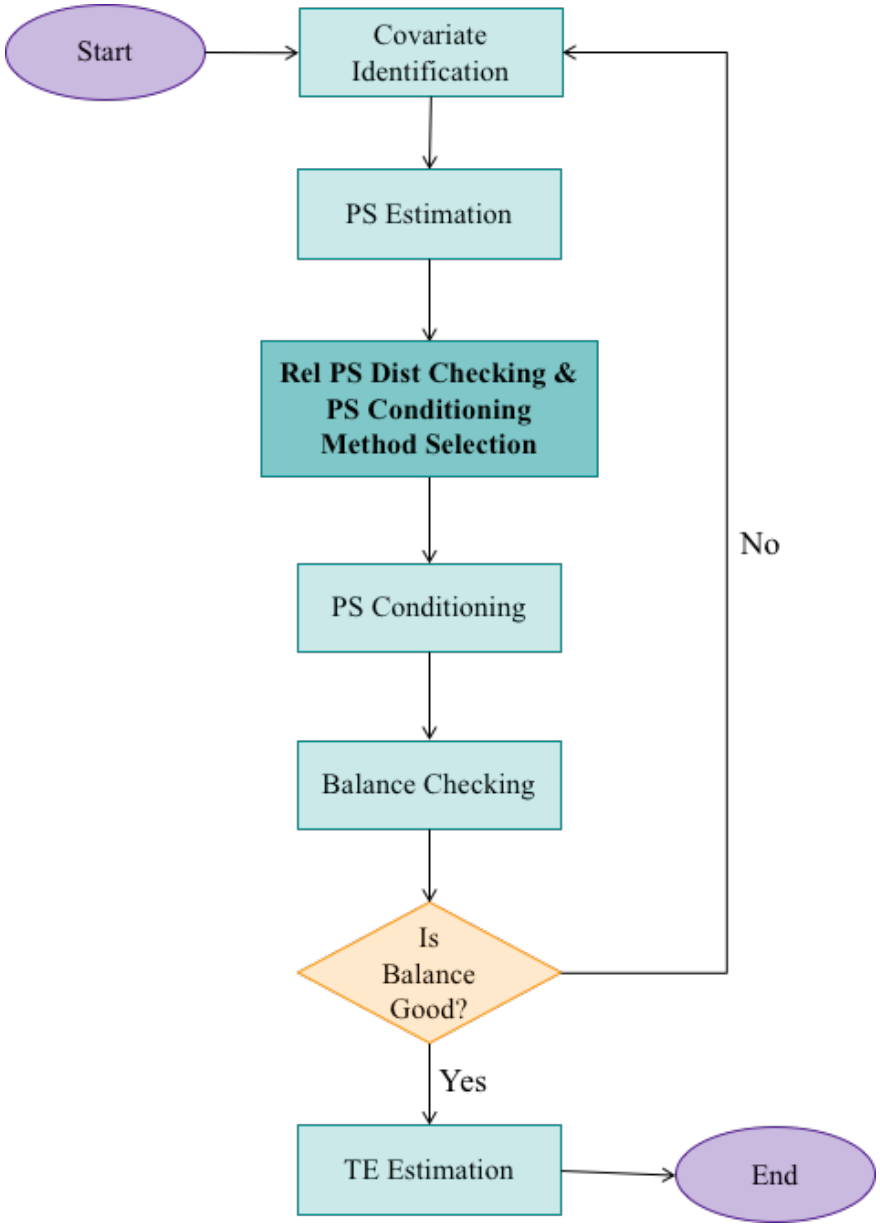


Figure 20. Flowchart of the new six-step PS procedure including relative PS distribution checking.

Appendix A: Balance Plot (Scenario B)

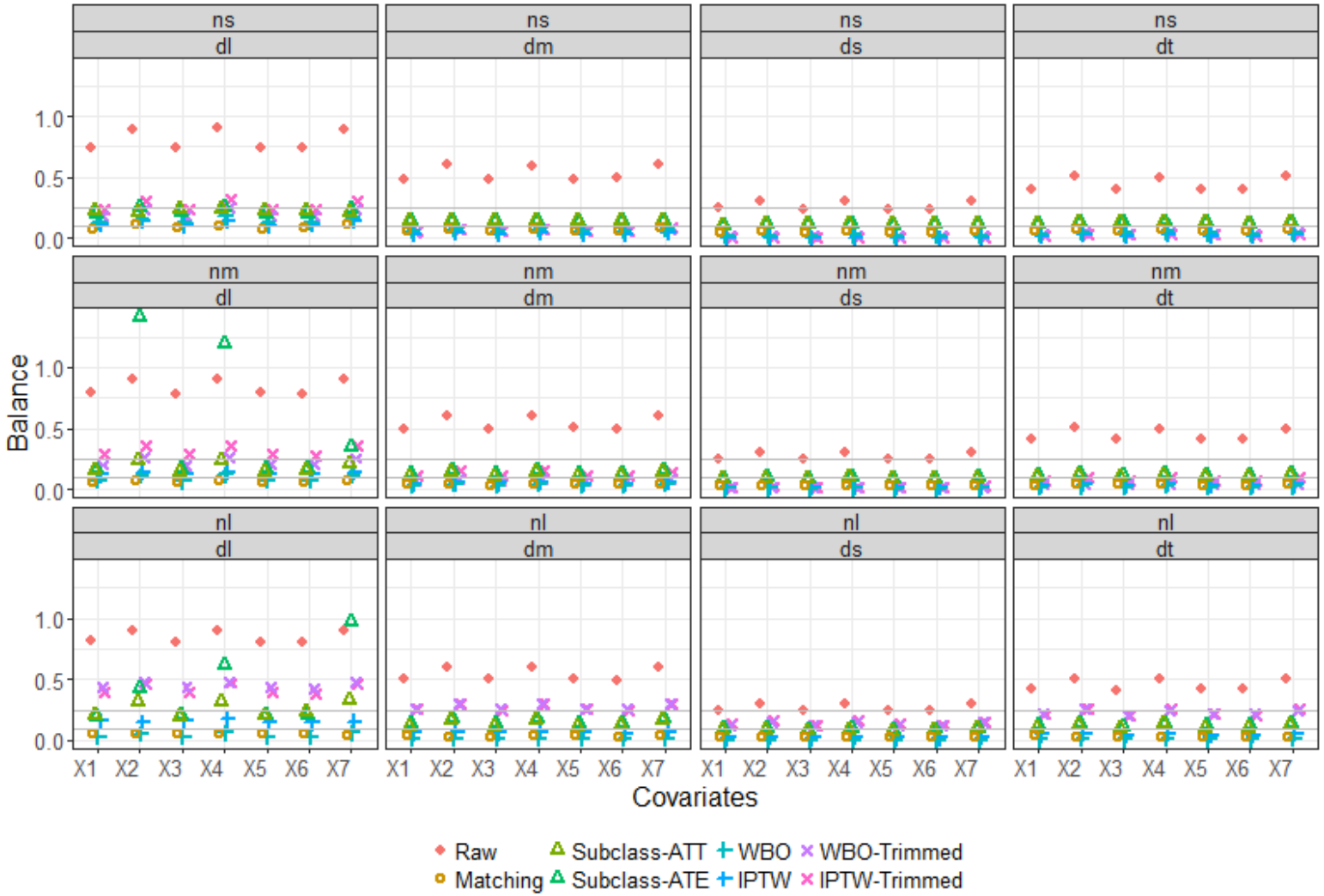


Figure A1. Standardized mean difference (SMD) for all PS conditioning models (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group.

Appendix B: Plots for Relative Bias in TE Estimates (Scenario B)

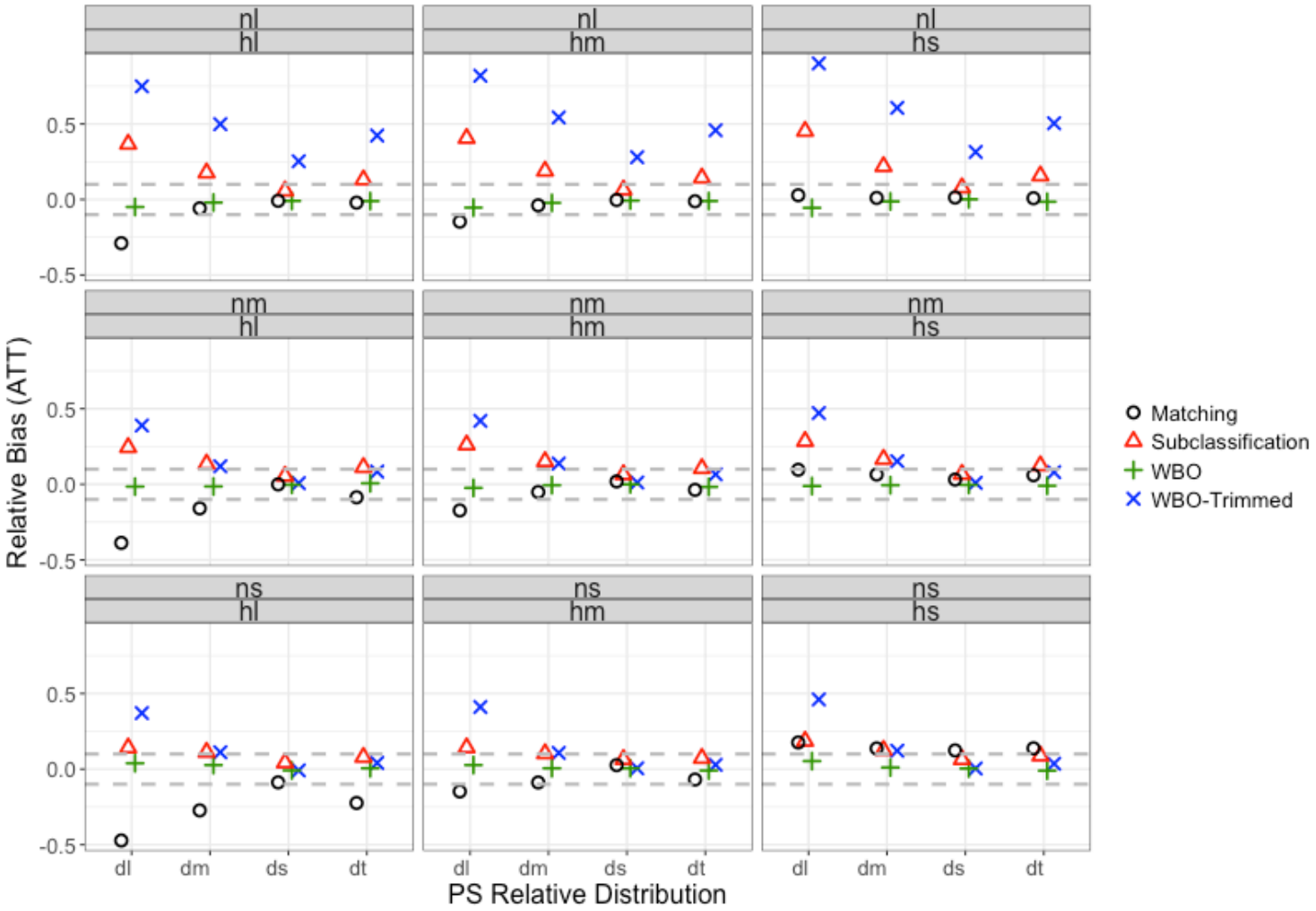


Figure B1. Relative bias for ATT methods (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in relative PSdistributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects, another way of saying substantial, slight, and zero heterogeneous treatment effects discussed in Chapter 3.

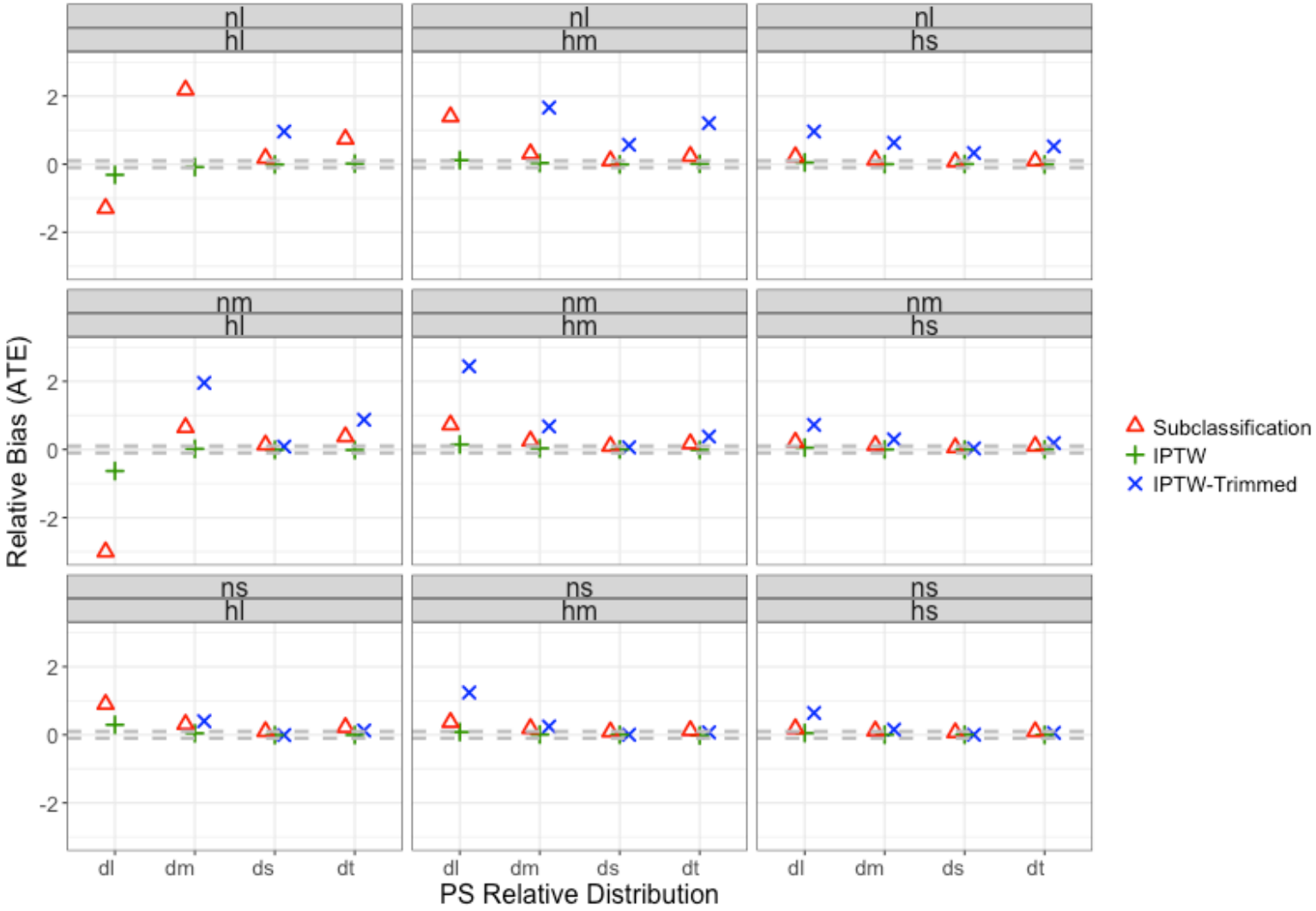


Figure B2. Relative bias for ATE methods (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects. Five large relative bias produced by IPTW with trimming are out of boundary of the plot. The values are -4.41, 11.75, 3.61, 5.06, -9.32, and 3.15 for nl_dl_hl, nl_dm_hl, nl_dt_hl, hl_dl_hm, nm_dl_hl, ns_dl_hl, respectively.

Appendix C: Plots for Empirical SEs (Scenario B)

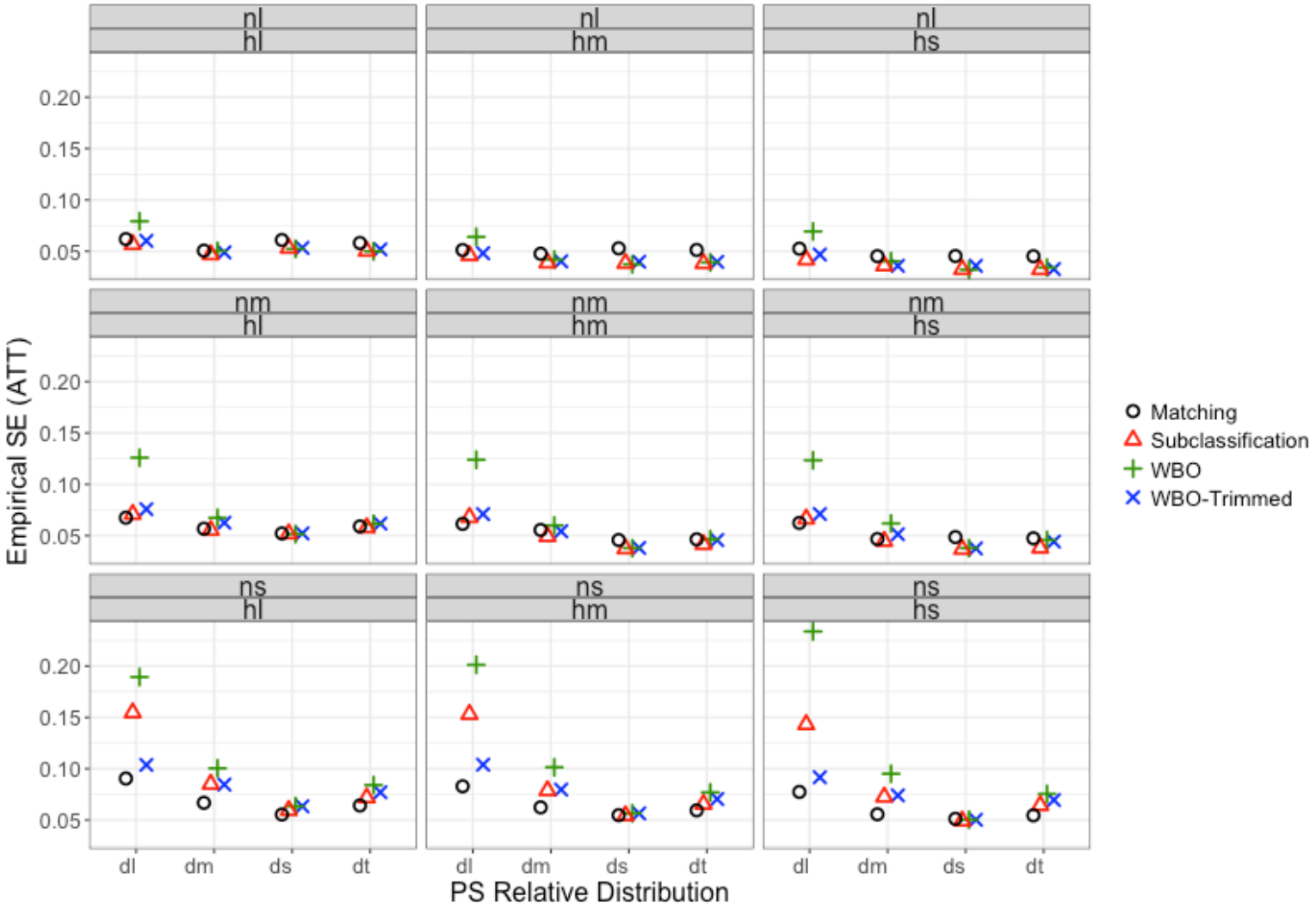


Figure C1. Empirical SE for ATT methods (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects.

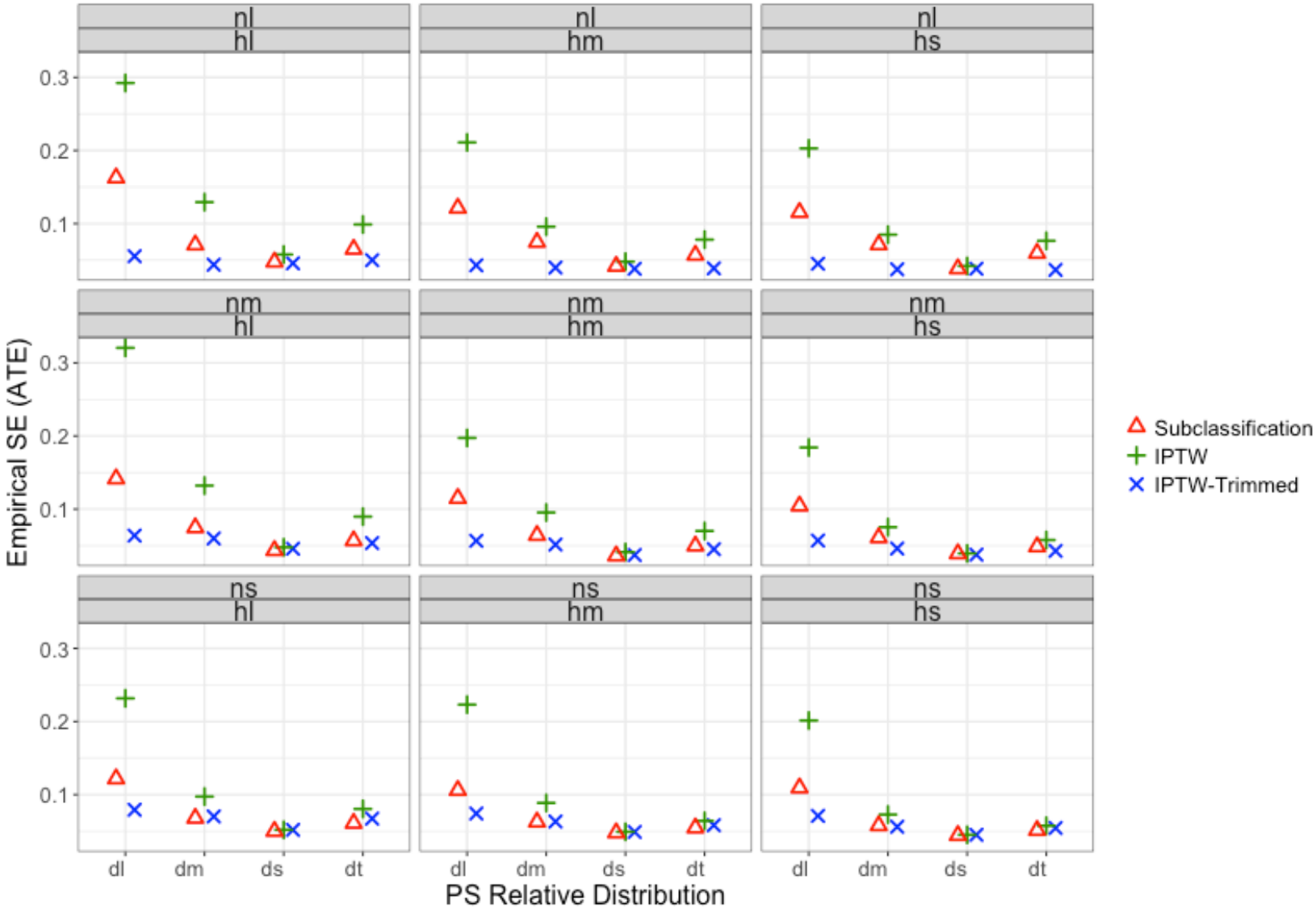


Figure C2. Standard Error for ATE methods (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects.

Appendix D: Plots for Relative Bias in SE Estimates (Scenario B)

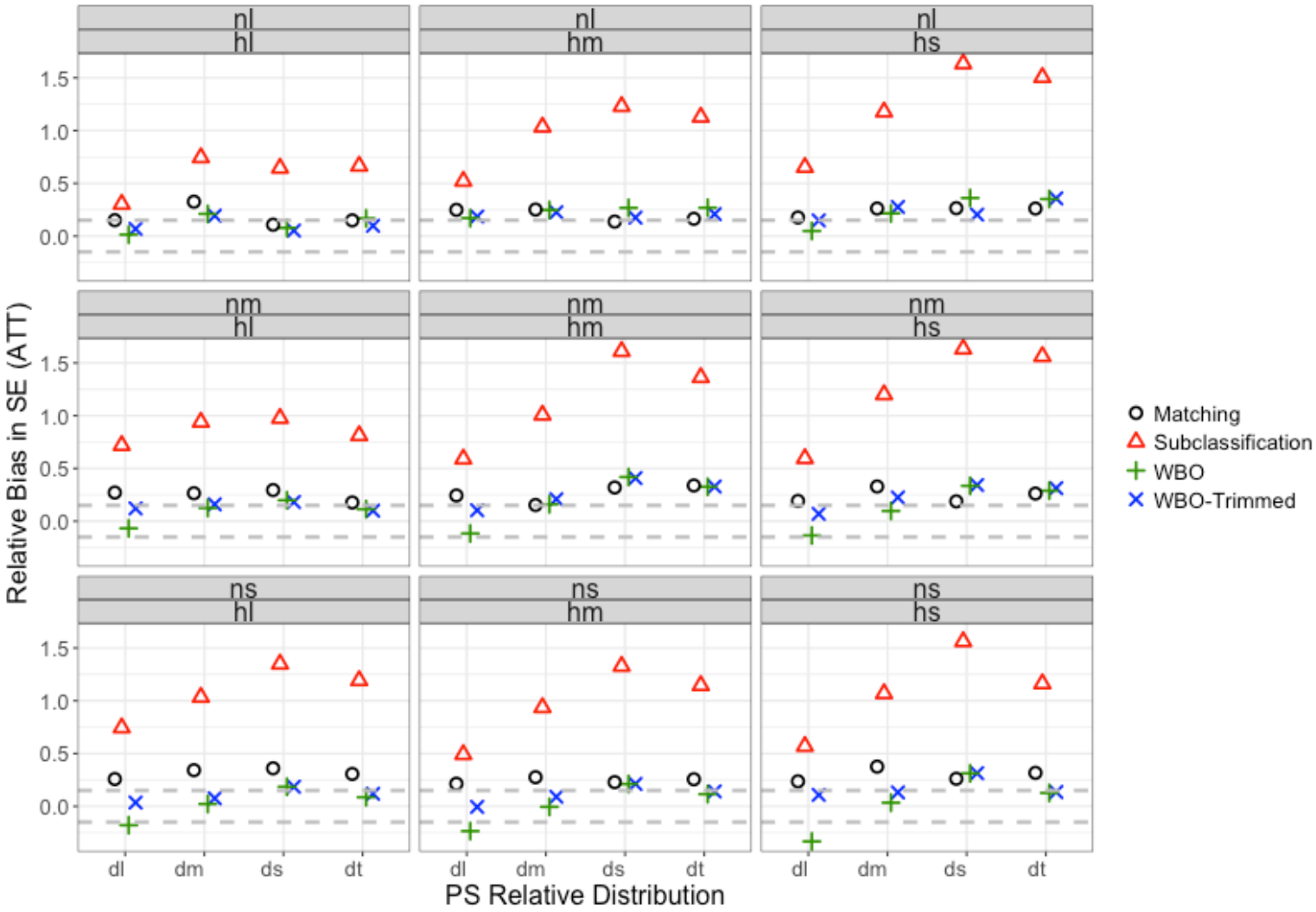


Figure D1. Relative bias in SE for ATT methods (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects. The grey dotted lines indicate the thresholds between *use with caution* and *untrustworthy* (±15%).

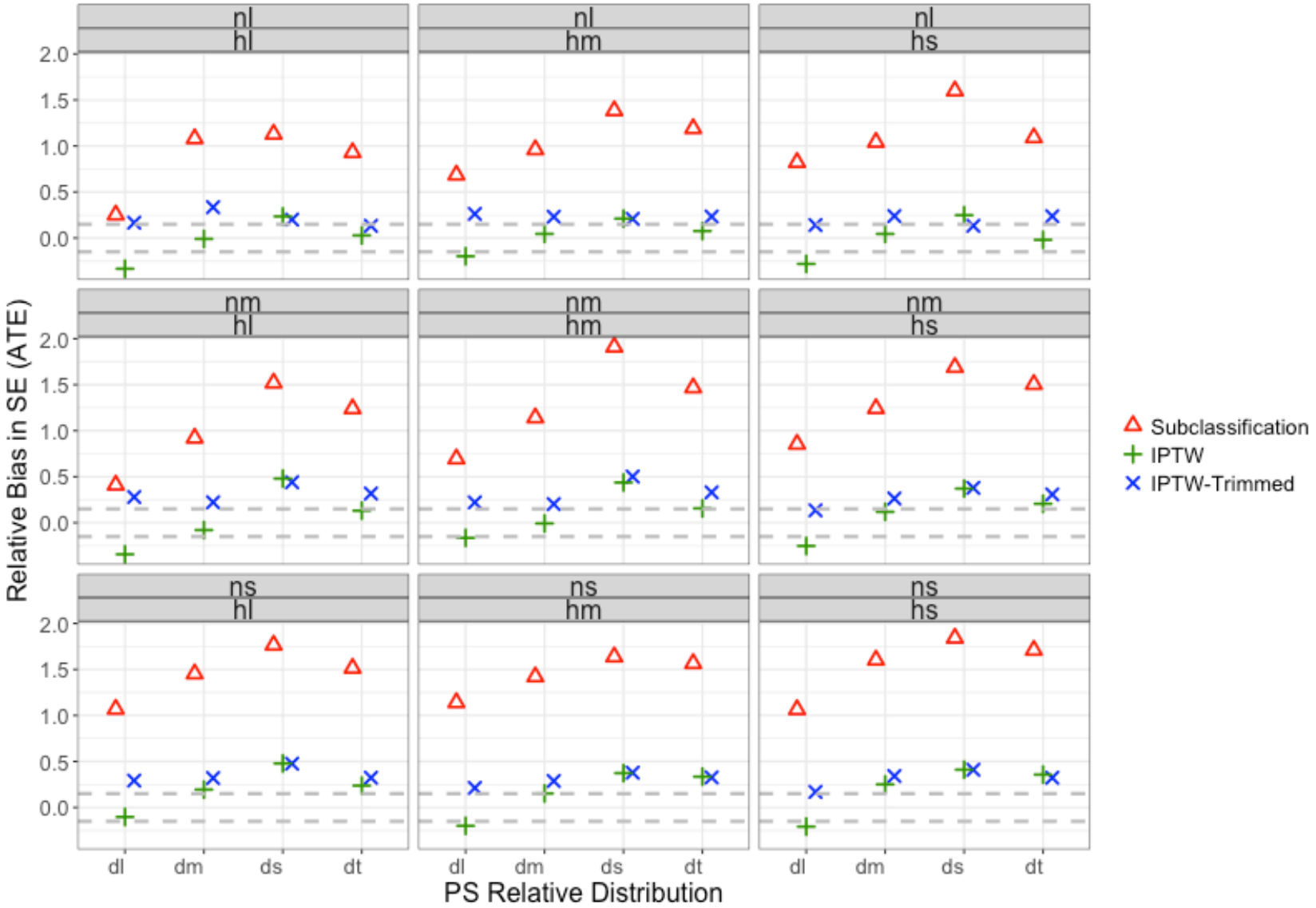


Figure D2. Relative bias in SE for ATE methods (Scenario B). dl(dm/ds/dt) = large(medium/small/truncated) difference in PS relative distributions. nl(nm/ns) = large(medium/small) sample size in the control group. hl(hm/hs) = large(medium/small) heterogeneous treatment effects. The grey dotted lines indicate the thresholds between *use with caution* and *untrustworthy* (±15%).

Appendix E: Tables for Scenario B

Table E1
Relative Bias in ATT Estimate

Conditions			Relative Bias (ATT)				
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Matching	Subclassification	WBO	WBO-Trimmed
Large	Truncated	Substantial	6.76	-0.02	0.13	-0.01	0.42
Large	Truncated	Slight	2.28	-0.01	0.14	-0.01	0.46
Large	Truncated	Zero	1.02	0.01	0.16	-0.02	0.50
Large	Large	Substantial	-7.80	-0.29	0.37	-0.05	0.75
Large	Large	Slight	9.15	-0.15	0.41	-0.05	0.82
Large	Large	Zero	1.75	0.03	0.45	-0.06	0.90
Large	Medium	Substantial	21.99	-0.06	0.18	-0.02	0.50
Large	Medium	Slight	3.15	-0.04	0.19	-0.02	0.54
Large	Medium	Zero	1.21	0.01	0.22	-0.01	0.61
Large	Small	Substantial	1.77	-0.01	0.06	-0.01	0.25
Large	Small	Slight	1.07	0.00	0.06	-0.01	0.28
Large	Small	Zero	0.62	0.01	0.08	0.00	0.31
Medium	Truncated	Substantial	4.06	-0.09	0.11	0.01	0.08
Medium	Truncated	Slight	1.93	-0.04	0.11	-0.02	0.07
Medium	Truncated	Zero	1.03	0.06	0.12	-0.01	0.08
Medium	Large	Substantial	-21.23	-0.39	0.24	-0.02	0.39
Medium	Large	Slight	5.81	-0.17	0.26	-0.02	0.42
Medium	Large	Zero	1.80	0.10	0.28	-0.01	0.47
Medium	Medium	Substantial	7.20	-0.16	0.14	-0.01	0.12
Medium	Medium	Slight	2.59	-0.05	0.15	-0.01	0.14
Medium	Medium	Zero	1.23	0.06	0.17	-0.01	0.15
Medium	Small	Substantial	1.42	0.00	0.05	0.00	0.01
Medium	Small	Slight	0.98	0.02	0.06	0.00	0.01
Medium	Small	Zero	0.62	0.03	0.07	0.00	0.01
Small	Truncated	Substantial	2.18	-0.22	0.08	0.01	0.04
Small	Truncated	Slight	1.49	-0.07	0.07	-0.01	0.03
Small	Truncated	Zero	1.03	0.14	0.09	-0.01	0.04
Small	Large	Substantial	8.86	-0.47	0.14	0.04	0.37
Small	Large	Slight	3.50	-0.15	0.14	0.03	0.41
Small	Large	Zero	1.83	0.18	0.19	0.05	0.46
Small	Medium	Substantial	3.05	-0.27	0.11	0.03	0.11
Small	Medium	Slight	1.91	-0.09	0.10	0.01	0.11
Small	Medium	Zero	1.24	0.14	0.13	0.01	0.12
Small	Small	Substantial	1.02	-0.09	0.04	-0.01	-0.01
Small	Small	Slight	0.82	0.03	0.06	0.01	0.01
Small	Small	Zero	0.62	0.12	0.06	0.00	0.00

Table E2

Relative Bias in ATE Estimate

Conditions			Relative Bias (ATE)			
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Subclassification	IPTW	IPTW-Trimmed
Large	Truncated	Substantial	6.76	0.74	0.02	3.61
Large	Truncated	Slight	2.28	0.23	0.01	1.20
Large	Truncated	Zero	1.02	0.10	-0.01	0.52
Large	Large	Substantial	-7.80	-1.29	-0.31	-4.41
Large	Large	Slight	9.15	1.40	0.12	5.06
Large	Large	Zero	1.75	0.22	0.05	0.96
Large	Medium	Substantial	21.99	2.18	-0.08	11.75
Large	Medium	Slight	3.15	0.31	0.03	1.66
Large	Medium	Zero	1.21	0.12	0.00	0.63
Large	Small	Substantial	1.77	0.18	-0.01	0.96
Large	Small	Slight	1.07	0.10	-0.01	0.57
Large	Small	Zero	0.62	0.07	0.00	0.33
Medium	Truncated	Substantial	4.06	0.38	-0.02	0.87
Medium	Truncated	Slight	1.93	0.17	-0.01	0.38
Medium	Truncated	Zero	1.03	0.10	0.00	0.19
Medium	Large	Substantial	-21.23	-3.00	-0.63	-9.32
Medium	Large	Slight	5.81	0.73	0.14	2.44
Medium	Large	Zero	1.80	0.22	0.05	0.72
Medium	Medium	Substantial	7.20	0.65	0.02	1.96
Medium	Medium	Slight	2.59	0.25	0.03	0.68
Medium	Medium	Zero	1.23	0.12	0.00	0.30
Medium	Small	Substantial	1.42	0.14	-0.01	0.08
Medium	Small	Slight	0.98	0.10	0.00	0.06
Medium	Small	Zero	0.62	0.06	-0.01	0.03
Small	Truncated	Substantial	2.18	0.22	-0.01	0.12
Small	Truncated	Slight	1.49	0.13	-0.02	0.07
Small	Truncated	Zero	1.03	0.09	0.00	0.05
Small	Large	Substantial	8.86	0.90	0.29	3.15
Small	Large	Slight	3.50	0.37	0.08	1.24
Small	Large	Zero	1.83	0.18	0.05	0.64
Small	Medium	Substantial	3.05	0.31	0.04	0.40
Small	Medium	Slight	1.91	0.19	0.01	0.24
Small	Medium	Zero	1.24	0.12	0.00	0.15
Small	Small	Substantial	1.02	0.10	-0.01	-0.01
Small	Small	Slight	0.82	0.09	0.00	0.00
Small	Small	Zero	0.62	0.06	0.00	0.00

Table E3
Empirical Standard Error for ATT Estimate

Conditions			Empirical Standard Error (ATT)				
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Matching	Subclassification	WBO	WBO-Trimmed
Large	Truncated	Substantial	0.06	0.06	0.05	0.05	0.05
Large	Truncated	Slight	0.04	0.05	0.04	0.04	0.04
Large	Truncated	Zero	0.04	0.05	0.03	0.03	0.03
Large	Large	Substantial	0.05	0.06	0.06	0.08	0.06
Large	Large	Slight	0.04	0.05	0.05	0.06	0.05
Large	Large	Zero	0.04	0.05	0.04	0.07	0.05
Large	Medium	Substantial	0.05	0.05	0.05	0.05	0.05
Large	Medium	Slight	0.04	0.05	0.04	0.04	0.04
Large	Medium	Zero	0.04	0.05	0.04	0.04	0.04
Large	Small	Substantial	0.06	0.06	0.05	0.05	0.05
Large	Small	Slight	0.05	0.05	0.04	0.04	0.04
Large	Small	Zero	0.04	0.05	0.03	0.03	0.04
Medium	Truncated	Substantial	0.06	0.06	0.06	0.06	0.06
Medium	Truncated	Slight	0.05	0.05	0.04	0.05	0.05
Medium	Truncated	Zero	0.04	0.05	0.04	0.05	0.04
Medium	Large	Substantial	0.05	0.07	0.07	0.13	0.08
Medium	Large	Slight	0.05	0.06	0.07	0.12	0.07
Medium	Large	Zero	0.05	0.06	0.07	0.12	0.07
Medium	Medium	Substantial	0.06	0.06	0.06	0.07	0.06
Medium	Medium	Slight	0.05	0.06	0.05	0.06	0.05
Medium	Medium	Zero	0.05	0.05	0.04	0.06	0.05
Medium	Small	Substantial	0.06	0.05	0.05	0.05	0.05
Medium	Small	Slight	0.05	0.05	0.04	0.04	0.04
Medium	Small	Zero	0.04	0.05	0.04	0.04	0.04
Small	Truncated	Substantial	0.06	0.06	0.07	0.08	0.08
Small	Truncated	Slight	0.06	0.06	0.07	0.08	0.07
Small	Truncated	Zero	0.05	0.05	0.06	0.08	0.07
Small	Large	Substantial	0.09	0.09	0.15	0.19	0.10
Small	Large	Slight	0.08	0.08	0.15	0.20	0.10
Small	Large	Zero	0.08	0.08	0.14	0.23	0.09
Small	Medium	Substantial	0.07	0.07	0.09	0.10	0.08
Small	Medium	Slight	0.06	0.06	0.08	0.10	0.08
Small	Medium	Zero	0.06	0.06	0.07	0.10	0.07
Small	Small	Substantial	0.06	0.06	0.06	0.06	0.06
Small	Small	Slight	0.05	0.05	0.05	0.06	0.06
Small	Small	Zero	0.05	0.05	0.05	0.05	0.05

Table E4

Empirical Standard Error for ATE Estimate

Conditions			Empirical Standard Error (ATE)			
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Subclassification	IPTW	IPTW-Trimmed
Large	Truncated	Substantial	0.06	0.07	0.10	0.05
Large	Truncated	Slight	0.04	0.06	0.08	0.04
Large	Truncated	Zero	0.04	0.06	0.08	0.04
Large	Large	Substantial	0.05	0.16	0.29	0.06
Large	Large	Slight	0.04	0.12	0.21	0.04
Large	Large	Zero	0.04	0.12	0.20	0.05
Large	Medium	Substantial	0.05	0.07	0.13	0.04
Large	Medium	Slight	0.04	0.07	0.10	0.04
Large	Medium	Zero	0.04	0.07	0.08	0.04
Large	Small	Substantial	0.06	0.05	0.06	0.05
Large	Small	Slight	0.05	0.04	0.05	0.04
Large	Small	Zero	0.04	0.04	0.04	0.04
Medium	Truncated	Substantial	0.06	0.06	0.09	0.05
Medium	Truncated	Slight	0.05	0.05	0.07	0.05
Medium	Truncated	Zero	0.04	0.05	0.06	0.04
Medium	Large	Substantial	0.05	0.14	0.32	0.06
Medium	Large	Slight	0.05	0.11	0.20	0.06
Medium	Large	Zero	0.05	0.10	0.18	0.06
Medium	Medium	Substantial	0.06	0.07	0.13	0.06
Medium	Medium	Slight	0.05	0.06	0.10	0.05
Medium	Medium	Zero	0.05	0.06	0.08	0.05
Medium	Small	Substantial	0.06	0.04	0.05	0.05
Medium	Small	Slight	0.05	0.04	0.04	0.04
Medium	Small	Zero	0.04	0.04	0.04	0.04
Small	Truncated	Substantial	0.07	0.06	0.08	0.07
Small	Truncated	Slight	0.06	0.05	0.06	0.06
Small	Truncated	Zero	0.06	0.05	0.06	0.05
Small	Large	Substantial	0.06	0.12	0.23	0.08
Small	Large	Slight	0.06	0.11	0.22	0.07
Small	Large	Zero	0.06	0.11	0.20	0.07
Small	Medium	Substantial	0.07	0.07	0.10	0.07
Small	Medium	Slight	0.06	0.06	0.09	0.06
Small	Medium	Zero	0.05	0.06	0.07	0.06
Small	Small	Substantial	0.07	0.05	0.05	0.05
Small	Small	Slight	0.06	0.05	0.05	0.05
Small	Small	Zero	0.06	0.04	0.05	0.05

Table E5

Relative Bias in the SE Estimate for ATT

Conditions			Relative Bias in SE Estimate (ATT)				
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Matching	Subclassification	WBO	WBO-Trimmed
Large	Truncated	Substantial	-0.22	0.15	0.66	0.17	0.10
Large	Truncated	Slight	-0.03	0.16	1.13	0.27	0.21
Large	Truncated	Zero	0.14	0.26	1.51	0.35	0.36
Large	Large	Substantial	-0.18	0.15	0.30	0.01	0.07
Large	Large	Slight	-0.01	0.25	0.52	0.17	0.19
Large	Large	Zero	0.02	0.18	0.65	0.05	0.15
Large	Medium	Substantial	-0.14	0.33	0.74	0.21	0.19
Large	Medium	Slight	-0.01	0.25	1.03	0.25	0.23
Large	Medium	Zero	0.17	0.26	1.18	0.22	0.28
Large	Small	Substantial	-0.24	0.11	0.65	0.08	0.05
Large	Small	Slight	-0.06	0.14	1.23	0.27	0.18
Large	Small	Zero	-0.02	0.26	1.63	0.36	0.20
Medium	Truncated	Substantial	-0.14	0.18	0.81	0.11	0.10
Medium	Truncated	Slight	-0.01	0.34	1.36	0.33	0.33
Medium	Truncated	Zero	0.11	0.26	1.56	0.29	0.31
Medium	Large	Substantial	-0.04	0.27	0.72	-0.07	0.12
Medium	Large	Slight	0.04	0.24	0.59	-0.12	0.10
Medium	Large	Zero	-0.05	0.19	0.60	-0.14	0.07
Medium	Medium	Substantial	-0.12	0.27	0.94	0.12	0.16
Medium	Medium	Slight	-0.04	0.15	1.01	0.16	0.21
Medium	Medium	Zero	0.03	0.33	1.20	0.10	0.22
Medium	Small	Substantial	-0.19	0.30	0.98	0.20	0.18
Medium	Small	Slight	0.02	0.32	1.61	0.42	0.41
Medium	Small	Zero	0.10	0.19	1.63	0.34	0.34
Small	Truncated	Substantial	-0.06	0.31	1.19	0.09	0.12
Small	Truncated	Slight	0.03	0.26	1.15	0.12	0.14
Small	Truncated	Zero	0.03	0.32	1.16	0.13	0.14
Small	Large	Substantial	0.05	0.26	0.74	-0.18	0.04
Small	Large	Slight	-0.07	0.21	0.49	-0.24	-0.01
Small	Large	Zero	0.02	0.24	0.57	-0.33	0.11
Small	Medium	Substantial	0.02	0.34	1.03	0.02	0.08
Small	Medium	Slight	-0.02	0.28	0.94	-0.01	0.09
Small	Medium	Zero	0.06	0.38	1.07	0.03	0.13
Small	Small	Substantial	0.03	0.36	1.35	0.19	0.19
Small	Small	Slight	-0.02	0.23	1.33	0.21	0.21
Small	Small	Zero	-0.03	0.26	1.56	0.31	0.31

Table E6

Relative Bias in the SE Estimate for ATE

Conditions			Relative Bias in SE Estimate (ATE)			
Sample Size	Rel. PS Dist.	Heter. TE	Naïve	Subclassification	IPTW	IPTW-Trimmed
Large	Truncated	Substantial	-0.22	0.93	0.03	0.13
Large	Truncated	Slight	-0.03	1.19	0.08	0.23
Large	Truncated	Zero	0.14	1.09	-0.02	0.24
Large	Large	Substantial	-0.18	0.25	-0.33	0.17
Large	Large	Slight	-0.01	0.69	-0.20	0.26
Large	Large	Zero	0.02	0.82	-0.28	0.14
Large	Medium	Substantial	-0.14	1.08	-0.01	0.33
Large	Medium	Slight	-0.01	0.96	0.04	0.23
Large	Medium	Zero	0.17	1.04	0.04	0.24
Large	Small	Substantial	-0.24	1.13	0.23	0.20
Large	Small	Slight	-0.06	1.38	0.21	0.21
Large	Small	Zero	-0.02	1.60	0.25	0.13
Medium	Truncated	Substantial	-0.14	1.24	0.13	0.32
Medium	Truncated	Slight	-0.01	1.47	0.16	0.33
Medium	Truncated	Zero	0.11	1.50	0.21	0.31
Medium	Large	Substantial	-0.04	0.41	-0.34	0.28
Medium	Large	Slight	0.04	0.70	-0.16	0.22
Medium	Large	Zero	-0.05	0.85	-0.25	0.14
Medium	Medium	Substantial	-0.12	0.92	-0.08	0.22
Medium	Medium	Slight	-0.04	1.14	-0.01	0.20
Medium	Medium	Zero	0.03	1.24	0.12	0.26
Medium	Small	Substantial	-0.19	1.52	0.48	0.44
Medium	Small	Slight	0.02	1.91	0.44	0.50
Medium	Small	Zero	0.10	1.69	0.37	0.38
Small	Truncated	Substantial	-0.06	1.51	0.24	0.32
Small	Truncated	Slight	0.03	1.57	0.33	0.33
Small	Truncated	Zero	0.03	1.71	0.36	0.32
Small	Large	Substantial	0.05	1.07	-0.10	0.29
Small	Large	Slight	-0.07	1.14	-0.20	0.21
Small	Large	Zero	0.02	1.06	-0.21	0.17
Small	Medium	Substantial	0.02	1.45	0.19	0.32
Small	Medium	Slight	-0.02	1.42	0.15	0.29
Small	Medium	Zero	0.06	1.60	0.25	0.34
Small	Small	Substantial	0.03	1.77	0.48	0.48
Small	Small	Slight	-0.02	1.64	0.37	0.38
Small	Small	Zero	-0.03	1.84	0.41	0.41

References

An, J. & Stapleton, L. M. (2015, April). Assessing the effect of home computers on math learning: application of a propensity score method for complex sample survey data. Paper presented at the annual meeting of the American Educational Research Association (AERA), SIG: Advanced Studies of National Databases, Chicago, IL, USA.

An, J. & Stapleton, L.M. (2016, May). *The use of nonparametric propensity score estimation with data obtained using a complex sampling design*. Paper presented at the Modern Modeling Methods (M3) Conference, Storrs, CT., USA.

Arpino, B., & Cannas, M. (2015). Comparing different approaches for propensity score matching with clustered data: a simulation study (RECSM Working Paper No. 43, 1-26). *Barcelona, Spain: University Pompeu Fabra*.

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

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

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

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, *33*, 1057-1069.

Becker H. J. (2000). Who's wired and who's not: children's access to and use of computer technology. *Future Child*, *10*, 44-75.

Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology*, *104*, 682.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Carroll, A. E., Rivara, F. P., Ebel, B., Zimmerman, F. J., & Christakis, D. A. (2005). Household computer and internet access: the digital divide in a pediatric clinic population. *AMIA Annual Symposium Proceedings*, *2005*, 111-115.

Cham, H. N. (2013). *Propensity score estimation with random forests* (Doctoral dissertation). Retrieved from Arizona State University.

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

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

Czajka, J. L., Hirabayashi, S. M., Little, R. J., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, *10*, 117-131.

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: applying propensity score methods to complex surveys. *Health Services Research*, *49*, 284-303.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*, 761-767.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, *589*, 63-93.

Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, *5*, 263-275.

Guo, F. (2009). Fairness of automated essay scoring of GMAT® AWA (GMAC® Research Reports: RR-09-01). *Retrieved from the Graduate Management Admission Council® website: http://www. gmac. com/NR/rdonlyres/FACE0811-B6F7-45A9-B57D-ED3703984B9A/0/RR0901_AWAFairness. Pdf*

Hahs-Vaughn, D. L. (2015). Investigating the use of propensity score analysis with complex samples. In W. Pan & H. Bai (Eds.), *Propensity score analysis: Fundamentals and developments*. New York, NY: Guilford.

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

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman Hall/CRC Press.

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A comparison of experimental and observational data analyses. In Gelman, A. & Meng, X. L. (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An*

essential journey with Donald Rubin's statistical family (pp. 49-60). New York, NY: John Wiley.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 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*, 199-236.

Hoffman, D. L., & Novak, T. P. (1998). Bridging the racial divide on the Internet. *Science*, *280*, 390-391.

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

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*, *101*, 901-910.

Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis*, *29*, 239-261.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: an application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, *44*, 407.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329-367.

Hullsiek, K. H., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, *3*, 179-193.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*, 1-24.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4-29.

Jiao, H., Zou, J., Liao, D., Li, C., & Lissitz, R. W. (2016, April). *A comparison of methods to link a state test to the PARCC consortium test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*, 81-97.

Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis*, *33*, 458-482.

Kishl, L. (1992). Weighting for unequal p_i. *Journal of Official Statistics*, *8*, 183-200.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under

conditions of nonuniform effect. *American Journal of Epidemiology*, *163*, 262-270.

Lechner, M., (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner and F. Pfeiffer (Eds.), *Econometric evaluations of active labor market policies in Europe* (pp. 43-58). Heidelberg, Physica.

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

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS ONE*, *6*, e18174.

Lenis, D., Ackerman, B., & Stuart, E. A. (2018). Measuring model misspecification: Application to propensity score methods with complex survey data. *Computational Statistics & Data Analysis*, *128*, 48-57.

Li, C., Liao, D., Zou J., Jiao, H., & Lissitz, R. W. (2016, June). *Investigating the concordance relationship between the MSA and PARCC scores using propensity score matching and extrapolation methods.* Paper presented at the Maryland State Department of Education Data Summit, Ellicott City, MD.

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*, 2937-2960.

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

Mojtabai, R., & Graff Zivin, J. (2003). Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis. *Health Services Research*, *38*, 233-259.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, *38*, 231-281.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431-462.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*, 317-337.

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods, American Statisticcal Association*, American Statistical Association, 225-230.

Robins, J. M., (1998). Marginal structural models. In *1997 Proceedings of the Section on Bayesian Statistical Science* (pp. 1-10). Alexandria, VA: American Statistical Association.

Robins, J. M., (1999). Marginal structural models versus structural nested modes as tools for causal inference. In E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology: The environment and clinical trials* (pp. 95-134). New York, NY: Springer-Verlag.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 550-560. Doi: 10.1097/00001648-200009000-00011. [PubMed: 10955408]

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

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

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, *41*, 103-116.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*, 159-184.

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

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, *52*, 249-264.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

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. *Pharmaco epidemiology and Drug Safety*, *17*, 546-555.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, *16*, 12-19.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: A review journal of the Institute of Mathematical Statistics*, *25*, 1-21.

Stuart, E.A. and Rubin, D.B. (2007). Best Practices in Quasi-Experimental Designs: Matching methods for causal inference. In Osborne, J. W. (Ed.), *Best Practices in Quantitative Social Science* (pp. 155-176). Thousand Oaks, CA: Sage Publications.

Stürmer, T., Wyss, R., Glynn, R. J., & Brookhart, M. A. (2014). Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *Journal of internal medicine*, *275*, 570-580.

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

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M. C., Leggitt, J., & Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten-First Grade Data File and Electronic Codebook, Public Version* (NCES 2015-078). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine-reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*, *357*, 2189-2194.

What Works Clearinghouse™ (2017). *WWC Standards Handbook Version 4.0*. Washington, DC: U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/handbooks.

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, *42*, 314-347.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, *4*, 67-91.

Zanutto, E., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics, 30*, 59-73.