Covariate Balance for Observational Effectiveness Studies: A Comparison of Matching and Weighting

Joseph M. Kush[1,2*], Elise T. Pas[2], Rashelle J. Musci[2], and Catherine P. Bradshaw[1]

[1]University of Virginia, School of Education and Human Development

[2]Johns Hopkins Bloomberg School of Public Health

Kush, J. M., Pas, E. T., Musci, R. J., & Bradshaw, C. P. (2022). Covariate balance for observational effectiveness studies: A comparison of matching and weighting. *Journal of Research on Educational Effectiveness*. http://dx.doi.org/10.1080/19345747.2022.2110545

Published in *Journal of Research on Educational Effectiveness*

*Correspondence concerning this article should be addressed to Joseph Kush, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway Room 841, Baltimore, MD 21205. Email: jkush1@jhu.edu

Acknowledgements: We thank the Maryland PBIS Management Team, which includes the Maryland State Department of Education, Sheppard Pratt Health System, and the 24 local school districts. We would also like to give special thanks to Drs. Ji Hoon Ryoo and Elizabeth Stuart for providing feedback on the paper and for their methodological consultation.

Funding: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H150027 and R305A150221 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Open Research Statements

Study and Analysis Plan Registration

- There is no registration associated with the study reported in this manuscript.

Data, Code, and Materials Transparency

- The code that support the findings of this study are openly available at: https://github.com/jmk7cj/Covariate-Balance

Design and Analysis Reporting Guidelines

- Not applicable.

Transparency Declaration

- The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

- This manuscript reports an original study.

Acknowledgements: We thank the Maryland PBIS Management Team, which includes the Maryland State Department of Education, Sheppard Pratt Health System, and the 24 local school districts. We would also like to give special thanks to Drs. Ji Hoon Ryoo and Elizabeth Stuart for providing feedback on the paper and for their methodological consultation.

Funding: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H150027 and R305A150221 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Propensity score matching and weighting methods are often used in observational effectiveness studies to reduce imbalance between treated and untreated groups on a set of potential confounders. However, much of the prior methodological literature on matching and weighting has yet to examine performance for scenarios with a majority of treated units, as is often encountered with programs and interventions that have been widely disseminated or "scaled up". Using a series of Monte Carlo simulations, we compare the performance of *k*:1 matching with replacement and weighting methods with respect to covariate balance, bias, and MSE. Results indicate that the accuracy of all methods declined as treatment prevalence increased. While weighting produced the largest reduction in covariate imbalance, 1:1 matching with replacement provided the most unbiased treatment effect estimates. An applied example using empirical school-level data is provided to further illustrate the application and interpretation of these methods to a real-world scale-up effort. We conclude by considering the implications of propensity score methods for observational effectiveness studies with a particular focus on educational research.

Keywords: propensity scores; matching; weighting; treatment prevalence

Experimentation through random assignment ensures that all variables related to both treatment receipt and the outcome (i.e., confounders) are balanced between treatment conditions. However, random assignment is not always feasible and is particularly unlikely to occur within observational research, frequently resulting in treated and untreated groups with unbalanced covariate distributions. Only after covariate balance between treatment groups is achieved may researchers obtain unbiased treatment effect estimates. Analytic approaches such as propensity score methods can be used to correct for such confounding. The propensity score can thus be viewed as a "balancing score", such that units with similar covariate distributions will have similar propensity scores, attempting to mimic random assignment (Rosenbaum & Rubin, 1983).

One condition often overlooked in the methodological literature on propensity scores is treatment prevalence, or the proportion of units exposed to treatment. Treatment prevalence is of great interest to researchers examining the effectiveness of interventions when widely disseminated or "brought to scale" because when effectively scaling an intervention, there is typically an increase in the proportion of units exposed to the intervention over time. Most notably, once more than 50% of a population (e.g., a school district, a state, or students) is exposed to treatment, matching with replacement becomes necessary, as failing to do so would result in the discarding of potentially important units, ultimately reducing statistical power in effect estimation (Stuart, 2010). When weighting for designs with a majority of treated units, certain untreated units may receive large or extreme weights if they comprise most of the information about the counterfactual, leading to increased variability in effect estimation (Austin & Stuart, 2015a; Hainmueller, 2011).

The current paper aims to fill this methodological gap in the educational literature by investigating the performance of propensity score matching and weighting methods with respect

to covariate balance for scenarios with varying degrees of treatment prevalence. This issue is of particular concern within the context of examining the effectiveness of a widely disseminated (or scaled-up) intervention, in which a majority of the sample exposed to treatment. More specifically, we build upon prior simulation research by Hainmueller (2011) and Colson et al. (2016) who compared inverse probability of treatment weighting and nearest neighbor matching with replacement for designs with treatment prevalence ranging from 0.2 to 0.5. By expanding our simulation to consider scenarios with a majority of units exposed to treatment, we broaden our understanding of propensity score weighting and matching methods to be more applicable to observational effectiveness designs. Additionally, we consider increasing degrees of imbalance in baseline covariates, representing scenarios in which there is an increasing difficulty in establishing covariate balance.

In fact, such methodological work has the potential to advance research on the effectiveness of educational interventions or programs already in wide use in “real-world” settings (see Fagan et al., 2019; Gottfredson et al., 2015). As compared to the amount of efficacy research in education and other fields such as public health or medicine, there is a paucity of research on the effectiveness of interventions. Such concerns in part motivated the U.S. Department of Education’s funding for what was initially coined “i3” (Investment in Innovation) and is currently the Education Innovation and Research funding (Office of Elementary and Secondary Education, 2021). Toward that end, we provide an applied example of a widely used and scaled educational preventive intervention framework, called Positive Behavioral Interventions and Supports, for which we use empirical data to illustrate the application and performance of different propensity score matching and weighting methods within a high treatment prevalence scenario.

Potential Outcomes Framework

Using the potential outcomes framework described by Neyman (1923) and Rubin (1974), consider the simplest case in which there are two groups or conditions (i.e., treatment and control), observed at a single time point. Let *Z* denote a binary treatment indicator. Specifically, each individual *i* is considered to have a potential outcome Y_i^1 associated with participating in the treatment condition ($Z_i = 1$), as well as a potential outcome Y_i^0 associated with participating in the non-treated (control) condition ($Z_i = 0$). The treatment effect for individual *i* is then defined as:

$$\delta_i = Y_i^1 - Y_i^0, \tag{1}$$

with the population ATE is given as:

$$\delta = E[Y^1 - Y^0]. \tag{2}$$

As only a single potential outcome for each individual is ever directly observed, several assumptions must be met for the expectations of the potential outcomes to be identified. One key assumption in observational studies is that treatment assignment is strongly ignorable. Rosenbaum and Rubin (1983) demonstrated that unbiased estimates of the ATE ($\hat{\delta}$) can be obtained if treatment assignment *Z* is independent of the potential outcome distribution of Y^1 and Y^0, conditional on an observed vector of covariates $\boldsymbol{X}$, and that no covariate values are associated with a probability of treatment equal to zero or one (positivity assumption). That is, $(Y^1, Y^0) \perp Z|\boldsymbol{X}$ and $0 < P(Z = 1|\boldsymbol{X}) < 1$. Additionally, the potential outcome distributions are also independent of treatment assignment *Z* given the propensity score $p(\boldsymbol{X})$, while treatment assignment *Z* is independent of the observed set of covariates $\boldsymbol{X}$, conditional on the propensity score:

$$(Y^1, Y^0) \perp Z|p(\boldsymbol{X}) \text{ and } Z \perp \boldsymbol{X}\,|p(\boldsymbol{X}). \tag{3}$$

The stable unit treatment value assumption (SUTVA; Rubin, 1986) is further assumed, requiring the potential outcomes of individual *i* to be independent of both the treatment assignment mechanism and the treatment status of other individuals.

To balance the treatment and control groups, probabilities of being in the treatment group are generated for all individuals. This probability is known as the propensity score, or the propensity of exposure to the treatment condition. The most common estimation function is a logistic regression model:

$$P(Z_i = 1|\boldsymbol{X}_i) = E(\boldsymbol{X}_i) = \frac{exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1 + exp(\boldsymbol{X}_i\boldsymbol{\beta})}. \quad (4)$$

where $\boldsymbol{X}_i$ represents a vector of observed covariates for individual *i*, while $\boldsymbol{\beta}$ represents a vector of associated parameter coefficients. After estimating propensity scores, individuals may be matched according to the similarity of scores; this is known as propensity score matching. Propensity scores may also be used as inverse probability weights in estimating the ATE, known as inverse probability of treatment weighting (Hernán & Robins, 2020). Both matching and weighting according to the propensity score have been shown to produce conditional independence between treatment assignment the outcome (Rosenbaum & Rubin, 1983), allowing for more accurate estimates of treatment effects.

In addition to estimating the ATE, propensity scores are commonly used to estimate the ATT. The ATT is defined as the difference in the expected value of the potential outcome for all individuals in the treatment group, had they been exposed to the control group ($\bar{Y}_i^0 \mid Z_i = 1$), from the expected value of the outcome for all individuals in the treatment group ($\bar{Y}_i^1 \mid Z_i = 1$; Heckman & Robb, 1984):

$$\delta = E[Y_i^1|Z_i = 1] - E[Y_i^0|Z_i = 1]. \quad (5)$$

Evaluation researchers and policymakers are often more substantively interested in the ATT than the ATE. In practice, school administrators may be interested in the treatment effect of a new behavioral intervention program for those who chose to participate in the program (ATT), not the effect for all students in the population (ATE).

Matching and Weighting Methods

Propensity score analyses provide multiple advantages over traditional regression-based analyses. For example, the propensity score is able to effectively summarize all covariate distributions in a single dimension, helping avoid the "curse of dimensionality" often encountered in regression-adjustment approaches. Additionally, propensity scores are estimated without reference to the eventual outcome of interest, often viewed as a more transparent method than regression adjustment (Greifer & Stuart, 2021). These properties make propensity score methods an attractive choice for researchers investigating effects in observational designs. Specific to the field of education, a number of more recent studies have explored the methodological implications of matching (Keele et al., 2020; Page et al., 2020; Pimentel et al., 2018; Rosenbaum, 2020) and weighting (Bishop et al., 2018; Fuentes et al., 2021; Leite et al., 2019) methods, demonstrating the utility and demand for robust causal inference methodology. In what follows, we describe each of these methods in greater detail, highlight advantages, disadvantages, and implications of certain method choices.

Matching

After estimating propensity scores, pairs of treated and untreated individuals may then be matched based according to the similarity of their propensity scores. One of the most common matching strategies is 1:1 nearest neighbor matching, in which a single untreated unit is matched with a single treated unit based on a distance measure (typically pairwise differences in

propensity scores between units). One key issue researchers face when conducting matching involves whether to match with or without replacement. When matching without replacement, after a single untreated unit is matched with a single treated unit, the untreated unit can no longer be paired with another treated unit. This has the potential to result in poor-quality matches for scenarios in which there are few untreated units. Oppositely, matching with replacement allows a single untreated unit to be matched with multiple treated units. Allowing for replacement may improve match quality if a given untreated unit is considered to be the best match (in terms of closest distance) for multiple treated units. Prior simulations by Austin (2013) demonstrated that matching with and without replacement both produce unbiased treatment effect estimates. However, when there is weak overlap in the covariate distributions, matching with replacement produces smaller standard errors.

One caveat to matching with replacement is that weights must be used in subsequent analyses (e.g., a weighted linear regression). The weights generated by matching with replacement reflect the frequency with which each untreated unit was matched (Leite, 2017). Let w_i represent the weight for unit *i*. Further, let w_i = 1 for all treated units (Z_i = 1). Then, weights for each untreated unit (Z_i = 0) are calculated as:

$$w_i = \frac{n_0}{n_1} * \sum_{m=1}^{n_i} \frac{1}{M_m}, \tag{6}$$

in which n_0 is the total number of matched cases, n_1 is the total number of treated cases, n_i is the number of treated cases unit *i* was matched to, and M_m is the total number of matches (including unit *i*) that each treated case received. Here, the untreated weights are scaled to sum to the total number of matched cases. Such weights are not utilized when conducting matching without replacement, as each untreated unit may only be matched with a single treated unit.

A second key issue involves *k*:1 matching, a matching strategy that finds and matches not one, but *k* untreated units to each treated unit. This may be particularly useful for scenarios with a majority of untreated units. For example, prior simulation research by Austin (2010) has demonstrated a bias-variance trade-off with increasing *k*. While increasing the number of untreated to treated matches tends to increase precision in treatment effect estimates, *k*:1 matching may simultaneously increase dissimilarity in the matched sample, ultimately increasing bias in effect estimates (Stuart, 2010). It is worth noting that while nearest neighbor matching is most commonly used in the health and social sciences (Thoemmes & Kim, 2011), other types of matching may provide certain benefits. For example, both Cepeda et al. (2003) and Ming and Rosenbaum (2000) found variable ratio matching, in which a variable number of untreated units are matched to each treated unit, resulted in better covariate balance but larger standard errors of the estimated treatment effect as compared to *k*:1 matching. Other types of matching, such as the optimal matching algorithm, entertain all possible matching combinations before declaring a match. While optimal matching has been shown to be remove more bias when allowing for a variable number of nearest neighbors (Gu & Rosenbaum, 1993), simulation studies have demonstrated nearest neighbor matching is more effective in reducing covariate imbalance than optimal matching when the untreated to treated ratio is fixed (Austin, 2013). Nonetheless, because nearest neighbor matching is the most commonly used matching approach, we focus on this particular matching strategy.

A third issue regards the use of a caliper or trimming. When using a caliper, no units further apart than some pre-specified distance measure are allowed to be matched. For example, if a distance measure of 0.2 standard deviations of the propensity score was chosen as caliper width, only those with propensity scores less than 0.2 standard deviations away from the unit of

interest could be considered to be matched with. Austin (2009a; 2011) and others have demonstrated that caliper widths of 0.6 and 0.2 standard deviations of the propensity score remove approximately 90% and 99% of bias due to measured confounders, respectively, by enforcing more similar matches. Calipers are similar but distinct from trimming, which refers to dropping units with propensity scores outside of a pre-specified range. This is sometimes referred to as the 'overlap problem' in which Crump et al. (2009) suggest discarding all units outside the range of (0.1, 0.9) standard deviations of the propensity score. However, one limitation to matching with a specified caliper width or trimming is that observations outside of the region of common support are often discarded, reducing the analytic sample size, potentially negatively impacting power (Ho et al., 2007a). Additionally, discarding observations outside of the region of common support shifts the estimand as the estimated treatment effect is no longer in regard to a full sample of treated units.

Weighting

Instead of matching on the propensity score, propensity scores can be used as unit weights when estimating the treatment effect. The intuition is that treated and untreated units are reweighted to be representative of the population of interest. The inverse probability of treatment weight (IPTW) is defined as $w_i = \frac{Z_i}{\hat{e}_i} + \frac{1-Z_i}{1-\hat{e}_i}$, where $\hat{e}_i$ is the estimated propensity score for unit i. Letting Y_i represent the observed outcome for unit i, the ATE may be estimated as (Austin & Stuart, 2015a):

$$\hat{\delta} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{Z_i Y_i}{\hat{e}_i}\right] - \left[\frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)Y_i}{1-\hat{e}_i}\right], \quad (7)$$

which represents the difference in the weighted average outcomes between the treated and untreated groups. To instead estimate the ATT, a new set of weights can be calculated as $w_i =$

$Z_i + (1 - Z_i)\frac{\hat{e}_i}{1-\hat{e}_i}$. Thus, while the weights used to estimate the ATE weight the treated and untreated groups to their respective populations, the weights used to estimate the ATT weight the untreated group only (to a representative population of the treatment group), as weights for those in the treated group are equal to one.

One concern with IPTW is that it can generate extreme or large weights. As the weights are directly related to the propensity scores, a (misspecified) propensity score model that produces extreme propensity scores may yield extreme weights. For example, when using weights to estimate the ATE, treated units with a propensity score close to zero may have very large weights. Similarly, when using weights to estimate the ATT, untreated units with a propensity score close to one may also have very large weights. Such large weights can potentially increase the standard error of the estimated treatment effect as well as increase bias (Harder et al., 2010; Leite, 2017). Thus, a misspecified propensity score model may derive extreme weights, although this may also be due to a lack of common support.

Solutions to address extreme weights include improving the specification of the propensity score model, and to conduct weight trimming or truncating (Lee et al., 2011). Trimming is typically performed by setting weights that exceed a specified threshold to that given threshold, often determined by quantiles of the distribution. For example, units with weights above the 95th percentile may be set equal to the 95th percentile (and vice versa for the 5th percentile). Prior research has demonstrated that weight trimming may decrease standard errors in treatment effect estimates, though weight trimming may also bias estimates depending on the estimation method (Lee et al., 2011; Thoemmes & Ong, 2016). However, there is no singular guideline for the optimal level of weight trimming. Thus, researchers conducting IPTW

must carefully inspect the distribution of weights and adjust either the propensity score model or the weights themselves.

Matching Versus Weighting

When matching with replacement, matching weights are ultimately used in outcome estimation to appropriately account for untreated units matched to multiple treated units. This begs the question, “If matching is a means to getting weights, why not weight directly?” The purpose of matching and weighting methods is the same, namely, to reduce bias in treatment effect estimates due to confounders for observational research. However, matching methods may provide an advantage in terms of robustness to model specifications. Generally, matching methods are less sensitive to correct specification in propensity score estimation than weighting methods (Waernbaum, 2012). As described earlier, extreme propensity scores may result in extreme weights, potentially resulting in effect estimates driven by a few units with large weights. For matching methods, the value of the propensity score itself is not directly used to compute matching weights. Additionally, matching methods allow for many possibilities to customize the matching procedure, such as deciding on a particular distance measure, the matching method to be used, whether or not to match with replacement, the number of untreated units to match with each treated unit, and the matching order, among others. Although this may result in a more cumbersome model building process than weighting, this has the advantages of decreasing bias and improving precision in effect estimation. Weighting can allow for a more efficient model building process, although researchers often customize the type of weighting (e.g., kernel, regularization, and entropy balancing weights). In general, weighting may be preferred to matching when: a) the form of the exposure model is known, or b) no units have extreme propensity scores. However, in practice the form of the exposure model is generally not

well known, favoring matching methods. See Greifer and Stuart (2021) for a more thorough comparison of matching and weighting methods.

Use of Propensity Score Methods in Observational Effectiveness Studies

As noted in the introduction, the current investigation of the performance of propensity score matching and weighting methods with respect to covariate balance for scenarios with varying degrees of treatment prevalence has the potential to inform observational effectiveness research on the "real-world" implementation of widely disseminated interventions; those effects may inherently differ from the experimental efficacy or even quasi-experimental study designs used in prior studies due to treatment prevalence differences. This process, in which the adoption of innovation varies throughout the course of an intervention results for the "diffusion of innovations" (Rogers, 1962), whereby the number of units implementing treatment would be expected to increase over time. Additionally, effectiveness research of a scaled intervention is likely to include relatively large proportions of the sample implementing treatment, such as more than 50% of units as treatment units (see Gottfedson et al., 2015; Fagan et al., 2019 for additional information on effectiveness research). In such instances of a scale-up, the number of untreated units is likely outnumbered by the number of treated units. As a result, matching with replacement becomes essential to ensure each treated unit is matched with at least one untreated unit. Additionally, the sensitivity of the weighting procedure to the sample specification may result in extreme weights.

Much of the previous simulation literature examining propensity score matching and weighting has only considered a relatively small proportion of treated units. For example, Austin (2009) only considered a constant 10% of units treated, whereas Austin (2010) considered a range from 2% treated to 15% treated. Moreover, because the proportion of treated units was so

small, both studies used matching without replacement. While Colson et al. (2016) considered a sample of 45% treated units for their simulations, this proportion is still smaller than those expected within observational effectiveness research of scaled-up interventions. Prior literature has yet to examine how propensity score matching and weighting perform for scenarios in which 50% or more of the sample is receiving treatment. This feature, common to observational effectiveness research of widely-disseminated interventions, suggests the utility of matching with replacement to retain the original sample. Taken together, the current paper extends previous simulation research on matching and weighting methods by comparing these methods for data with a range of treatment prevalence rates while also matching with replacement.

The Present Study

We investigated the performance of *k*:1 matching with replacement and IPTW across increasing levels of treatment prevalence. As the performance of these propensity score methods has yet to be examined for scenarios with a majority of treated units, the results from our study help inform observational research designs more broadly, with specific applicability toward observational effectiveness studies of widely disseminated interventions. Building upon prior simulation research, we conducted a series of Monte Carlo simulations to examine the performance of these two propensity score methods across a variety of scenarios faced by applied researchers. We build upon prior simulation research by Hainmueller (2011) and Colson et al. (2016) who compared inverse probability of treatment weighting and nearest neighbor matching with replacement for designs with treatment prevalence ranging from 0.2 to 0.5. By expanding our simulation to consider scenarios with a majority of units exposed to treatment, we broaden our understanding of propensity score weighting and matching methods to be more applicable to intervention scale-up research. Additionally, we consider increasing degrees of

imbalance in baseline covariates, representing scenarios in which there is an increasing difficulty in establishing covariate balance. An applied example using empirical data from a widely disseminated educational framework implemented in all states in the U.S. is further provided to illustrate the application and performance of propensity score matching and weighting methods for observational effectiveness studies with a high treatment prevalence.

Method

Design of the Simulation

To investigate the effect of the proportion of sample exposed to treatment we simulated data where the following design factors were manipulated: treatment prevalence, baseline imbalance, sample size, number of covariates, and propensity score method. The values and methods chosen for the simulation conditions are informed by previous propensity score simulation research and prior applied education research; they are meant to represent a broad range of designs researchers face in practice. Each of the manipulated factors are described below.

Treatment Prevalence

Previous simulation research on propensity score methods has not examined samples with a majority of units exposed to treatment. As described earlier, a large prevalence of treated units in the sample is a scenario often encountered in observational scale-up designs (see Gottfredson et al., 2015). However, prior simulation research has explored ratios of treated to untreated units less than or equal to 0.55 (Austin, 2013; Colson et al., 2016; Leite et al., 2019). As such, values of treatment prevalence were manipulated to $P = 0.2, 0.4, 0.6$, and 0.8. Importantly, the two conditions with a majority of treated units have not been explored in prior

propensity score literature and are most relevant for applied researchers using propensity score methods for observational effectiveness studies of scaled interventions.

Baseline Imbalance

The degree of baseline imbalance represents the standardized mean difference in covariate values for the treated and untreated groups. For continuous variables, the standardized mean difference (*d*) is defined as:

$$d = \frac{(\bar{X}_t - \bar{X}_c)}{\sqrt{\frac{{s_t}^2 + {s_c}^2}{2}}}, \quad (8)$$

where $\bar{X}_t$ and $\bar{X}_c$ represent the mean of the covariate for the treatment and control groups, respectively, while ${s_t}^2$ and ${s_c}^2$ represent the variance of the covariate for the treatment and control groups, respectively. Standardized mean differences larger than 0.10 have been suggested as representing meaningful imbalance (Austin & Mamdani, 2006). Accordingly, standardized mean difference values at baseline were manipulated to *d* = 0.2, 0.3, and 0.5 For two populations of equal size, a standardized mean difference of *d* = 0.2 yields approximately 85% overlap between the distributions; *d* = 0.3 yields approximately 79% overlap; and *d* = 0.5 yields 67% overlap (Cohen, 1988). Thus, the larger the standardized mean difference value in baseline covariates, the stronger the separation between the treated and untreated distributions (i.e., less overlap), representing a scenario that is increasingly difficult for a given propensity score method to achieve balance.

Sample Size

Sample sizes were varied to be *N* = 250 and 500, representing a small to medium sample size commonly found in school-based and student-based studies (Bradshaw et al., 2021; Lee & Gage, 2020). Moreover, these values are similar to or smaller than those used in prior propensity

score simulations (Stuart et al., 2013; Austin & Stuart, 2015b; Whittaker, 2020), thus representing a lower end of sample sizes.

Number of Covariates

The number of baseline covariates were set to X = 10 and 20. A common number of covariates included in previous propensity score matching and weighting studies is 10 (Austin & Stuart, 2015; Leite et al., 2019). While examples of researchers using propensity score methods to balance more than 50 covariates may be more common in the medical literature (see Austin et al., 2020), it may not be feasible to include a large set of covariates, particularly in smaller samples (Stuart, 2010).

Propensity Score Method

Finally, we compare propensity score matching and weighting methods by examining 1:1 matching with replacement, 3:1 matching with replacement, 5:1 matching with replacement, and ATT weighting. For matching, *k*:1 greedy nearest neighbor matching with replacement was used to match cases on the estimated propensity score. The greedy algorithm matches treatment units with control units without considering a global distance measure (e.g., Mahalanobis distance; Gu & Rosenbaum, 1993). No caliper was used in the matching process to allow for a retention of the full sample for analyses without changing the quantity of interest (Ho, Imai, King & Stuart, 2007b). For weighting, probability weights were calculated as $w_i = Z_i + (1 - Z_i)\frac{\hat{e}_i}{1-\hat{e}_i}$. Thus, we refer to this as ATT weighting. For both matching and weighting, a parametric logistic regression, in which all covariates were linearly related to a binary treatment variable, was used to estimate propensity scores. In all cases, the ATT was the estimand of interest. This resulted in a total of $4 \times 3 \times 2 \times 2 \times 4 = 192$ unique simulation cells. A total of 1,000 datasets were generated and analyzed according to each condition.

Data Generation

For each subject, either ten or twenty baseline covariates were drawn from independent standard normal distributions. The treatment-selection model for the ten-covariate scenario was given as:

$$\begin{aligned}\text{logit}[\Pr(Z=1)] = \alpha_0 + \alpha_1 X_1 + \alpha_1 X_2 + \alpha_1 X_3 + \alpha_1 X_4 + \alpha_1 X_5 \\ + \alpha_1 X_6 + \alpha_1 X_7 + \alpha_1 X_4 X_5 + \alpha_1 X_1 X_1 + \alpha_1 X_7 X_7,\end{aligned} \quad (9)$$

and for the twenty-covariate scenario as:

$$\begin{aligned}\text{logit}[\Pr(Z=1)] = \alpha_0 + \alpha_1 X_1 + \alpha_1 X_2 + \alpha_1 X_3 + \alpha_1 X_4 + \alpha_1 X_5 + \alpha_1 X_6 \\ + \alpha_1 X_7 + \alpha_1 X_8 + \alpha_1 X_9 + \alpha_1 X_{10} + \alpha_1 X_{11} + \alpha_1 X_{12} + \alpha_1 X_{13} + \alpha_1 X_{14} \\ + \alpha_1 X_9 X_{10} + \alpha_1 X_7 X_8 + \alpha_1 X_1 X_1 + \alpha_1 X_{10} X_{10} + \alpha_1 X_3 X_3 + \alpha_1 X_7 X_7.\end{aligned} \quad (10)$$

Thus, in both scenarios the true propensity score model included non-additivity and non-linearity, mimicking the complexity involved in real-world data. To determine the correct intercept value α_0 corresponding to a desired ratio of treated to untreated units, a bisection approach was used in which numerous intercept values are attempted until a given value results in the desired ratio within some pre-specified tolerance level; the specific algorithm is provided on lines 103-148 of the R code available at: https://github.com/jmk7cj/Covariate-Balance. The regression coefficient $\alpha_1$was equal to the standardized mean difference as shown in Equation 8. Then, for each subject, a binary treatment indicator *Z* was generated from a binomial distribution with a probability of exposure equal to $\frac{e^x}{1+e^x}$, where *x* is equal to the logit probability of treatment defined in Equations 9 and 10.

Next, potential outcomes for the ten-covariate scenario were generated as:

$$\begin{aligned}Y = \beta_0 + \delta_Z + \beta_1 X_1 + \beta_1 X_2 + \beta_1 X_3 + \beta_1 X_4 + \beta_1 X_5 \\ + \beta_1 X_8 + \beta_1 X_9 + \beta_1 X_2 X_4 + \beta_1 X_3 X_5 + \beta_1 X_1 X_1 + \varepsilon_i,\end{aligned} \quad (11)$$

and for the twenty-covariate scenario as:

$$Y = \beta_0 + \delta_Z + \beta_1 X_1 + \beta_1 X_2 + \beta_1 X_3 + \beta_1 X_4 + \beta_1 X_5 + \beta_1 X_6 + \beta_1 X_7 + \beta_1 X_8 + \beta_1 X_9 + \beta_1 X_{10} + \beta_1 X_{15} + \beta_1 X_{16} + \beta_1 X_{17} + \beta_1 X_{18} + \beta_1 X_3 X_5 + \beta_1 X_1 X_1 + \varepsilon_i. \quad (12)$$

Here, β_0 represents the intercept, β_1 represents the regression coefficient, δ_Z represents the population treatment effect on the treated, and ε_i represents a residual error term. The intercept value of β_0 was fixed to 0. The treatment effect δ was set to 0.5, indicating a medium effect size. The residual error term was normally distributed $\varepsilon_i \sim N(0,1)$. The regression coefficient was varied such that the R^2 of the covariates on the outcome was equal to 0.3. Overall, these values are in line with those found in meta-analyses of educational outcomes (Hedges & Hedberg, 2014) and prior simulation studies (Leite et al., 2019). Figures 1 and 2 provide a graphical depiction of the data generation process for the ten and twenty covariate scenarios, in which half of the covariates were true confounders, with other variables related to treatment only, the outcome only, or neither treatment nor the outcome.

Evaluation Criteria

The quality of the matching or weighting procedure was evaluated through balance diagnostics. Steiner et al. (2010) offered a rule of thumb for sufficiently good balance for covariates with standardized mean difference values less than $|0.10|$. We focus on the standardized mean difference averaged across the covariates after matching or weighting.

After estimating propensity scores, a subsequent weighted linear regression model was fit to the matched or weighted data using the survey package (Lumley, 2004), in which the ATT was estimated and compared to the known population treatment effect on the treated (δ). We considered unadjusted estimates (i.e., a single *treatment* independent variable). Absolute bias in ATT estimates was used to evaluate the model, with values given on the scale of Y^0. Finally, the performance of each method was further evaluated using mean squared error (MSE = bias^2 +

variance), in which the variance of ATT estimates across replications is used. There is no single cutoff value for which to determine adequate MSE, though values closer to zero indicate no bias, and no variability in estimates across replications. All analyses were conducted in R (R Core Team, 2021). Matching was implemented using the MatchIt package (Ho et al., 2007), while weighting was implemented using the WeightIt package (Greifer, 2019). All code used for data generation and analyses is available at: https://github.com/jmk7cj/Covariate-Balance

Results

Results with respect to the three measures of evaluation criteria (covariate imbalance, bias, and MSE) were largely similar across the two sample size conditions (N = 250 and 500). Likewise, results were largely similar across the two covariate conditions (X = 10 and 20). Generally, covariate imbalance, bias, and MSE decreased as sample size increased, while covariate imbalance, bias, and MSE increased as the number of covariates increased, although these effects were negligible. We therefore do not present all possible variations, but rather focus our discussion on conditions with N = 500 and X = 10 as illustrations.

Covariate imbalance, bias, and MSE are displayed in Figure 3 whereby each column within the plot represents a scenario with a specific baseline imbalance (e.g., d = 0.2, 0.4, and 0.6 “standard mean difference”). Standardized mean differences averaged across covariates after implementing propensity score methods are presented in the top panel of Figure 3. The middle panel provides results for absolute bias in ATT estimates. The bottom panel illustrates values of MSE across the replications. For each panel, the x-axis represents treatment prevalence (P = 0.2, 0.4, 0.6, and 0.8). Overall, it can be seen that as baseline imbalance increases, covariate imbalance, bias, and MSE increase, regardless of the propensity score method or treatment prevalence.

Covariate Imbalance

Focusing on covariate imbalance (top panel), it can be seen that ATT weighting outperformed *k*:1 matching with replacement in terms of reduction in covariate imbalance across a range of scenarios. For scenarios with a baseline imbalance of 0.2 standardized mean difference (averaged across the covariates), all propensity score methods achieved sufficient balance (SMD less than 0.1), although weighting resulted in the largest reduction in imbalance. For scenarios with larger imbalance at baseline, all methods performed worse. Notably, as treatment prevalence increased, it became increasingly difficult to achieve sufficient balance. While weighting resulted in the best balance on average, there were differences in performance within matching methods. As *k*:1 matching increased from 1:1 to 5:1, there were greater reductions in covariate imbalance. Particularly for scenarios with 80% treatment prevalence, 5:1 matching performed similarly to ATT weighting. We note that assessing balance on higher order moments such as the variance ratio is also important to ensure comparable groups. We do not report such findings here as balance with respect to variance ratios was largely similar across all simulations.

Bias

Next, we focus on bias in unadjusted ATT estimates (middle panel). Interestingly, results appear somewhat opposite to those found regarding reduction in covariate imbalance. While weighting resulted in the best balance across conditions, weighting also produced the largest bias in ATT estimates across conditions. This result is somewhat perplexing at first glance, as improved balance should be directly related to the removal of bias in effect estimates. However, simulation studies by Lee et al. (2010) and Stuart et al. (2012) have demonstrated that balance across covariates does not align perfectly with balance on the propensity score itself, and

ultimately recommend against assessing balance on the propensity score but rather on covariates. For scenarios with larger imbalance at baseline, all propensity score methods resulted in larger bias. For matching methods specifically, bias increased as *k*:1 matching increased from 1:1 to 5:1. There was no meaningful difference in bias across treatment prevalence.

MSE

Finally, we direct our attention to MSE of unadjusted ATT estimates (bottom panel). There is no clear or consistent best method in terms of smallest MSE for scenarios with an average covariate baseline imbalance of $d = 0.2$ or 0.4. However, with an average baseline imbalance of 0.6 standardized mean difference, weighting produces the largest MSE across levels of treatment prevalence. Among the matching methods, the choice of *k*:1 matching that resulted in the smallest MSE depended on the treatment prevalence. For example, 1:1 matching has the largest MSE of the matching methods when treatment prevalence equals 0.2, but the smallest MSE when treatment prevalence equals 0.6. In general, MSE increased as baseline imbalance increased regardless of the propensity score method, results similar to those found for covariate imbalance and bias outcomes.

For Applied Researchers: Illustrative Example

We now consider a case example using empirical Positive Behavioral Interventions and Supports (PBIS) administrative data from the state of Maryland to illustrate the application of *k*:1 matching with replacement and weighting methods within the context of a real-world high treatment prevalence scenario. School-level data from 1,316 K-12 public schools across the state involved in the state-wide scale-up were utilized. Data from the 2007-08 through the 2012-13 school year were provided by the Maryland State Department of Education. Demographic information included variables such as student enrollment, the percent of students receiving free

and reduced-price meals, and the suspension rate. The outcome of interest was the truancy rate (i.e., percent of students missing 20 or more days of school in a school year) in a given year, as prior literature has demonstrated evidence that PBIS may reduce school-level truancy rates (Bradshaw et al., 2021; Pas et al., 2019). PBIS implementation data was also collected each year, in which schools implementing PBIS were considered treated units, while all other schools were considered untreated units. Because treatment assignment was not random, selection bias between treated and untreated schools was accounted for using either matching or weighting methods.

We focused on PBIS implementation and outcome data during two separate periods: the 2007-08 school year and 2013-14 school year. We focused on these two years to highlight a common theme in intervention scale-up study designs, wherein the proportion of the sample implementing treatment grows over time. Approximately 38% of schools implemented PBIS during the 2007-08 school year, while approximately 66% of schools implemented PBIS during the 2013-14 school year. For each timepoint, the ATT effect of PBIS on truancy rates was estimated.

To estimate propensity scores, demographic variables from the current year were included as predictors of current year PBIS status. Specifically, student enrollment, the percent of students receiving free and reduced-price meals, the percent of students who were African American, the suspension rate, and the percent of students who were proficient or advanced on the state standardized tests of reading and math were used as predictors of PBIS status. These variables were included based on previous research by Pas et al. (2019), with an overall aim to reduce selection bias between treated and untreated schools.

Following the procedures outlined in the simulation study, propensity scores were estimated using a parametric logistic regression with all covariates linearly related to a binary PBIS indicator variable with the ATT as the estimand of interest. After estimating propensity scores, matching and weighting were conducted, including 1:1, 3:1, and 5:1 nearest neighbor matching with replacement, as well as ATT weighting. Standardized mean differences for each covariate before and after matching or weighting were calculated. Finally, a weighted linear regression model was fit to the data, in which unadjusted estimates of the effect of PBIS on truancy were estimated:

$$\text{Truancy}_i = \beta_0 + \beta_1 \text{PBIS}_i + \varepsilon_i \quad (13)$$

Here, Truancy_i is the percentage of students missing 20 or more days of school in a school year for school *i*, β_0 is the intercept, $\beta_1$is the treatment effect of PBIS, and ε_i is a residual error term.

Table 1 provides standardized mean differences before and after matching or weighting. During the 2007-08 school year, in which approximately 38% of schools were implementing PBIS, the average standardized mean difference of the covariates at baseline was $d = 0.15$. Overall, 5:1 matching resulted in the largest reduction in covariate imbalance, reducing standardized mean differences by approximately 76% to an average of $d = 0.04$. While weighting reduced standardized mean differences by approximately 58% to an average of $d = 0.07$ (within the 0.1 rule of thumb), both 3:1 and 5:1 matching outperformed weighting. See Figure 4 for a graphical depiction of covariate balance before and after matching or weighting.

During the 2013-14 school year, in which approximately 66% of schools implemented PBIS, the average standardized mean difference of the covariates at baseline was $d = 0.35$. This represents a more difficult scenario to achieve balance than the 2007-08 school year, as treatment prevalence and the degree of imbalance at baseline were both larger. Similar to the earlier

timepoint, 5:1 matching resulted in the largest reduction in covariate imbalance, reducing standardized mean differences by approximately 92% to an average of $d = 0.03$. Weighting performed worst, although still reducing standardized mean differences by approximately 75% to an average of $d = 0.09$, still below the acceptable cutoff. Figure 5 provides a graphical depiction of covariate balance before and after matching or weighting.

Discussion

The current article sought to compare propensity score matching and weighting methods when used to reduce imbalance between treated and untreated groups on a set of potential confounders, with particular interest in the situation where the number of treated units exceeds the number of untreated. Under such conditions, propensity score matching with replacement becomes necessary to ensure sample size considerations. This was the first study to examine propensity score matching and weighting methods relevant to address this issue, and has particular significance in the field of effectiveness research on widely disseminated (or "scaled") interventions. As such, this study has potentially important implications for the design of policy-relevant research designs which aim to determine the impact of programs and interventions being brought to scale with implementation in a majority of relevant settings (i.e., schools).

Simulation results across a wide range of scenarios demonstrated that as treatment prevalence increases to greater than 50%, both 1:1 matching with replacement and ATT weighting perform worse in terms of reducing covariate imbalance, bias in ATT estimates, and the MSE of ATT estimates as compared to 3:1 and 5:1 matching with replacement. While no single method resulted in the smallest MSE across conditions, the results were more consistent for covariate imbalance and bias. Regarding reductions in baseline imbalance across covariates, ATT weighting produced the largest reductions of all propensity score methods, regardless of the

degree of baseline imbalance or treatment prevalence. As *k*:1 matching increased from 1:1 to 5:1, the covariates achieved greater balance, only equaling the imbalance resulting from weighting when treatment prevalence was 80%. Thus, weighting was the superior method for reducing covariate imbalance.

On the other hand, weighting also produced the largest bias in ATT estimates, regardless of the degree of imbalance or treatment prevalence. As previously described, this finding may appear counterintuitive but is consistent with prior simulation work that demonstrated a distinction between reducing imbalance among covariates and reducing imbalance on the propensity score itself (Lee et al., 2010; Stuart et al., 2012). For covariates with nonnormal distributions or scenarios in which some covariates are not related to treatment assignment, the differences between establishing balance on covariates versus the propensity score may grow larger. This is in line with the data generation process used in the current simulations, in which only 20% of the covariates were related to the outcome. As a result, although weighting resulted in the largest reductions in covariate imbalance, weighting also produced the largest bias in treatment effect estimates. Regarding bias, 1:1 matching with replacement was the superior method for producing unbiased treatment effect estimates.

Propensity score matching and weighting methods share the same goal; to reduce bias in treatment effect estimates due to confounders in observational research. Therefore, assessing covariate balance can ultimately be viewed as a means to an end, in which methods that reduce imbalance generally lead to unbiased effect estimates. Our simulation results demonstrated that while weighting achieved the greatest covariate balance, 1:1 matching ultimately produced the most unbiased effect estimates. This result speaks to the traditional bias-variance trade-off, in which increasing *k* tends to decrease the standard error of treatment effect estimates but also

increase bias in treatment effect point estimates (Austin, 2010). One must decide then which property to emphasize: precise but biased estimates or unbiased but imprecise estimates. Our results demonstrate that this trade-off may be optimized by increasing to 3:1 or 5:1 matching with replacement. We therefore recommend slightly increasing the number of nearest neighbors used (i.e., more than one) for researchers examining the effectiveness of widely disseminated interventions being brought to scale in observational studies. While matching with replacement involves extra steps by the analyst as compared to weighting methods, these come with the benefit of producing less biased effect estimates. Moreover, the use of matching with replacement is particularly important for such research designs, as the majority of units may be exposed to treatment, requiring the reuse of control units in the matching process. While matching with replacement still results in the use of matching weights incorporated into outcome analyses, the benefits of not relying upon weights that are directly related to propensity scores can be seen through reductions in bias.

Our results demonstrated that performance deteriorated (i.e., less reduction in covariate imbalance, larger bias in ATT estimates, larger MSE of ATT estimates) as treatment prevalence increased, regardless of sample size, the number of covariates, baseline imbalance, or the propensity score method of choice. This increase in treatment prevalence across time is a defining feature of widely disseminated interventions, as demonstrated through our empirical example in which PBIS implementation rates increased from 38% to 66% over time. The PBIS example depicted here is one clear example of widespread dissemination both within this state and nationally. This is not unique to PBIS; currently, social emotional learning curricula and restorative justice/practices in schools are also seeing widespread national dissemination. Even programs that are not widely disseminated nationally may be of interest, given widespread local

dissemination (e.g., within large districts or within states). Programs that are scaled-up may have efficacy research supporting their use or that efficacy research may come after or in tandem with dissemination. In either case, interventions that are widely disseminated would benefit from real-world effectiveness studies. In conducting such studies, researchers must be cognizant that the same matching and weighting methods that achieved well balanced groups to allow for unbiased effect estimates at lower treatment prevalence rates may not work as well for larger treatment prevalence.

In fact, the simulation results demonstrated that both low and high treatment prevalence often negatively impact findings. This is demonstrated in Figure 3 with nonlinearity in the graphs. Additional sensitivity analyses not shown here demonstrated that optimal performance often occurred when treatment prevalence was evenly distributed at 50%. This again yields interesting implications for observational effectiveness researchers. Practically speaking, the ability to estimate unbiased treatment effects with low treatment prevalence would be most beneficial for promising interventions, and fewer resources would be required in implementation. In juxtaposition, it may become irrelevant and even impossible to estimate treatment effects as treatment prevalence approaches 100% as there is no comparison control group. Thus, scenarios with a balanced ratio of treated and untreated units may provide the most accurate estimates. Taken together, these findings have import implications for researchers examining the effects of programs or interventions being brought to scale in observational settings, an area of limited research but great relevance. In particular, we urge researchers to be cognizant of such sampling design issues from the onset, recognizing that investigations of treatment effects in the very early stages (e.g., 20% treatment prevalence) or the very late stages (e.g., 80% treatment prevalence) of implementation roll out may complicate one’s ability to produce unbiased effect estimates.

Study Limitations

There are important limitations to our simulation study that should be considered. First, propensity scores were estimated using a linear combination of covariates in a parametric logistic regression, a much more simplistic model than the known data generative process. As the functional form of the true propensity score model is rarely known in practice, machine learning methods such as classification and regression trees or random forest models may provide flexibility in propensity score estimation (e.g., Lee et al., 2010; McCaffrey et al., 2004; Suk & Kang, 2021). Similarly, nonparametric weighting methods such as marginal mean weighting through stratification (MMWS; Hong, 2010; 2012), a combination of nonparametric estimators with entropy balancing (Vegetabile et al., 2021), and covariate balancing generalized propensity score methods (CBGPS; Fong et al., 2018) may provide alternatives when the functional form of the exposure model is unknown. Additionally, we considered unadjusted treatment effect estimates, in which a binary treatment indicator was the sole predictor in the outcome model. Doubly robust estimators may provide advantages so long as either the propensity score model or the outcome model are correctly specified (Nguyen et al., 2017). Thus, including covariates in a more complex outcome model may improve the accuracy and precision of treatment effect estimates. A final, critical limitation concerns the longitudinal nature of scale-up study designs. While the current paper conducted propensity score analyses cross-sectionally, a more complex approach may consider time-varying treatment. Marginal structural models may be used to appropriately handle time-dependent confounding, estimating the probability of treatment at each timepoint, independent of prior covariate and treatment histories.

Conclusions

Despite these limitations, our study provides useful information for observational researchers examining the effects of interventions being brought to scale. This paper illustrates how covariate balance and treatment effect estimates are impacted by treatment prevalence. Notably, a larger prevalence of treated units is associated with greater imbalance and larger bias in effect estimates. Findings from our series of studies suggest that *k*:1 matching with replacement results in less biased ATT estimates than propensity score weighting across a range of treatment prevalence rates, and that increasing *k* to larger than one may optimize such bias-variance trade-offs. We recommend researchers consider the findings of this paper when planning and designing quasi-experimental study designs, particularly when examining interventions that have been widely scaled-up.

References

Austin, P. C. (2009a). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal, 51*(1), 171-184. https://doi.org/10.1002/bimj.200810488

Austin, P. C. (2009b). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083-3107. https://doi.org/10.1002/sim.3697

Austin, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *Practice of Epidemiology, 172*(9), 1092-1097. https://doi.org/10.1093/aje/kwq224

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10*(2), 150-161. https://doi.org/10.1002/pst.433

Austin, P. C. (2013). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33*(6), 1057-1069. https://doi.org/10.1002/sim.6004

Austin, P. C., Wu, C. F., Lee, D. S., & Tu, J. V. (2020). Comparing the high-dimensional propensity score for use with administrative data with propensity scores derived from high-quality clinical data. *Statistical Methods in Medical Research, 29*(2), 568-588. https://doi.org/10.1177/0962280219842362

Austin, P. C., & Stuart, E. A. (2015a). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661-3679. https://doi.org/10.1002/sim.6607

Austin, P. C., & Stuart, E. A. (2015b). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research, 26*(4), 1654-1670. https://doi.org/10.1177/0962280215584401

Bishop, C. D., Leite, W. L., & Snyder, P. (2018). Using propensity score weighting to reduce selection bias in large-scale data sets. *Journal of Early Intervention, 40*(4), 347-362. https://doi.org/10.1177/1053815118793430

Bradshaw, C. P., Pas, E. T., Musci, R. J., Kush, J. M., & Ryoo, J. H. (2021). Can policy promote adoption or outcomes of evidence-based prevention programming?: A case illustration of Positive Behavioral Interventions and Supports. *Prevention Science, 22*(7), 986-1000. https://doi.org/10.1007/s11121-021-01257-0

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Optimal matching with a variable number of controls vs. a fixed number of controls for a cohort study: Trade-offs. *Journal of Clinical Epidemiology, 56*(3), 230-237. https://doi.org/10.1016/S0895-4356(02)00583-8

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

Colson, K. E., Rudolph, K. E., Zimmerman, S. C., Goin, D. G., Stuart, E. A., van der Laan, M., & Ahern, J. (2016). Optimizing matching and analysis combinations for estimating causal effects. *Scientific Reports, 6,* 23222. https://doi.org/10.1038/srep23222

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika, 96*(1), 187-199. https://doi.org/10.1093/biomet/asn055

Fagan, A. A., Bumbarger, B. K., Barth, R. P., Bradshaw, C. P., Cooper, B. R., Supplee, L. H., & Walker, D. K. (2019). Scaling up evidence-based interventions in US public systems to prevent behavioral health problems: Challenges and opportunities. *Prevention Science, 20*(8), 1147–1168. https://doi.org/10.1007/s11121-019-01048-8

Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics, 12*(1), 156-177. https://doi.org/10.1214/17-AOAS1101

Fuentes, A., Lüdtke, O., & Robitzsch, A. (2021). Causal inference with multilevel data: A comparison of different propensity score weighting approaches. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2021.1925521

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science, 16*(7), 893-926. https://doi.org/10.1007/s11121-015-0555-x

Greifer, N. (2019). WeighIt: Weighting for covariate balance in observational studies. R package version 0.12.0. https://CRAN.R-project.org/package=WeightIt

Greifer, N. (2021). cobalt: Covariate balance tables and plots. R package version 4.3.1. https://CRAN.R-project.org/package=cobalt

Greifer, N., & Stuart, E. A. (2021). Matching methods for confounders adjustment: An addition to the epidemiologist's toolbox. *Epidemiological Reviews*. https://doi.org/10.1093/epirev/mxab003

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405-420. https://doi.org/10.2307/1390693

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*(1), 25-46. https://doi.org/10.1093/pan/mpr025

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234–249. https://doi.org/10.1037/a0019623

Heckman, J., & Robb, R. (1984). Alternative methods for evaluating the impact of interventions. In J. Heckman and B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 156-245). Cambridge University Press.

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445-489. https://doi.org/10.1177/0193841X14529126

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1-28. https://doi.org/10.18637/jss.v042.i08

Hong, G. (2010) Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics, 35*(5), 499-531. https://doi.org/10.3102/1076998609359785

Hong, G. (2012) Marginal mean weighting through stratification: a generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods, 17*(1), 44-60. https://doi.org/10.1037/a0024918

Lee, A., & Gage, N. A. (2020). Updating and expanding systematic reviews and meta-analyses on the effects of school-wide positive behavior interventions and supports. *Psychology in the Schools, 57*(5), 783-804. https://doi.org/10.1002/pits.22336

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346. https://doi.org/10.1002/sim.3782

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one, 6*(3), e18174. https://doi.org/10.1371/journal.pone.0018174

Leite, W. (2017). *Practical propensity score methods using R.* Thousand Oakes, CA: Sage Publications.

Leite, W. L., Aydin, B., & Gurel, S. (2019). A comparison of propensity score weighting methods for evaluating the effects of programs with multiple versions. *The Journal of Experimental Education, 87*(1), 75-88. https://doi.org/10.1080/00220973.2017.1409179

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9*(8), 1-19. https://doi.org/10.18637/jss.v009.i08

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*(1), 118-124. https://doi.org/10.1111/j.0006-341X.2000.00118.x

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. Translated in *Statistical Science, 5*, 465-480, 1990.

Nguyen, T., Collins, G. S., Spence, J., Daurès, J., Devereaux, P. J., Landais, P., & Manach, Y. L. (2017). Double-adjustment in propensity score matching analysis: Choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology, 17*(78). https://doi.org/10.1186/s12874-017-0338-0

Office of Elementary and Secondary Education. (2021, June 30). *Education innovation and research*. https://oese.ed.gov/offices/office-of-discretionary-grants-support-services/innovation-early-learning/education-innovation-and-research-eir/

Page, L. C., Lenard, M. A., & Keele, L. (2020). The design of clustered observational studies in education. *AERA Open, 6*(3), 1-14. https://doi.org/10.1177/2332858420954401

Pas, E. T., Ryoo, J. H., Musci, R., & Bradshaw, C. P. (2019). The effects of a state-wide scale-up of school-wide Positive Behavior Intervention and Supports on behavioral and academic outcomes: A quasi-experimental examination. *Journal of School Psychology, 73*, 41-55. https://doi.org/10.1016/j.jsp.2019.03.001

Pimentel, S. D., Page, L. C., Lenard, M., & Keele, L. J. (2018). Optimal multilevel matching using network flows: An application to a summer reading intervention. *Annals of Applied Statistics, 12*(3), 1479-1505. https://doi.org/10.1214/17-AOAS1118

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria. https://www.R-project.org/

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application, 7*, 143-176. https://doi.org/10.1146/annurev-statistics-031219-041058

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1986). Comment: Which ifs have causal answers? *Journal of the American Statistical Association, 81*, 961-962. https://doi.org/10.1080/01621459.1986.10478355

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250–267. https://doi.org/10.1037/a0018719

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1-21. https://doi.org/10.1214/09-STS313

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research.

Journal of Clinical Epidemiology, 66(8), S84-S90.e1. https://doi.org/10.1016/j.jclinepi.2013.01.013

Suk, Y., Kang, H. (2021). Robust machine learning for treatment effects in multilevel observational studies under cluster-level unmeasured confounding. *Psychometrika*. https://doi.org/10.1007/s11336-021-09805-x

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118. https://doi.org/10.1080/00273171.2011.540475

Thoemmes, F., & Ong, A. D. (2016). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood, 4*(1), 40-59. https://doi.org/10.1177/2167696815621645

Vegetabile, B. G., Griffin, B. A., Coffman, D. L., Cefalu, M., Robbins, M. W., & McCaffrey, D. F. (2021). Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology, 21*(1), 69-110. https://doi.org/10.1007/s10742-020-00236-2

Waernbaum, I. (2012). Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. *Statistics in Medicine, 31*(15), 1572-1581. https://doi.org/10.1002/sim.4496

Whittaker, T. A. (2020). The comparison of latent variable propensity score models to traditional propensity score models under conditions of covariate unreliability. *Multivariate Behavioral Research, 55*(4), 625-646. https://doi.org/10.1080/00273171.2019.1663136

Table 1

Standardized Mean Differences Before and After Matching or Weighting

	2007-08 School Year (PBIS prevalence = 38%)								
	Baseline	Weighting		1:1 matching		3:1 matching		5:1 matching	
	d	*d*	%	*d*	%	*d*	%	*d*	%
Enrollment	-0.03	0.00	-95.1	0.09	162.1	0.08	132.2	0.05	42.3
FARMs	0.13	-0.04	-67.9	-0.13	0.3	-0.10	-19.0	-0.07	-48.2
% AA	-0.01	-0.04	623.6	-0.03	370.5	-0.02	196.2	-0.01	42.3
% Suspend	0.31	-0.11	-65.1	0.05	-83.6	0.04	-88.2	0.04	-88.4
Math	-0.25	0.10	-61.2	0.09	-65.6	0.01	-94.4	0.02	-91.9
Read	-0.19	0.10	-49.6	0.12	-35.5	0.05	-74.4	0.04	-80.6
Average	**0.15**	**0.07**	**-57.8**	**0.09**	**-44.8**	**0.05**	**-67.2**	**0.04**	**-76.4**
	2012-13 School Year (PBIS prevalence 66%)								
	Baseline	Weighting		1:1 matching		3:1 matching		5:1 matching	
	d	*d*	%	*d*	%	*d*	%	*d*	%
Enrollment	0.02	0.04	61.8	0.09	296.6	0.00	-88.3	0.02	7.3
FARMs	0.50	-0.07	-85.7	-0.11	-79.0	-0.04	-91.3	-0.04	-92.6
% AA	0.26	-0.10	-60.6	-0.15	-44.3	-0.06	-77.5	-0.05	-80.4
% Suspend	0.46	-0.16	-65.7	0.04	-90.7	0.03	-93.2	0.04	-91.9
Math	-0.44	0.07	-84.0	0.06	-87.3	-0.01	-98.3	-0.01	-98.7
Read	-0.42	0.08	-80.3	0.04	-91.4	0.01	-98.7	0.01	-98.7
Average	**0.35**	**0.09**	**-75.1**	**0.08**	**-77.4**	**0.02**	**-92.9**	**0.03**	**-92.3**

Note: Weighting = ATT weighting; 1:1 matching = 1:1 nearest neighbor matching with replacement, etc. *d* = standardized mean difference; % = percentage reduction in absolute value standardized mean difference, in which negative values indicate a decrease in *d*.

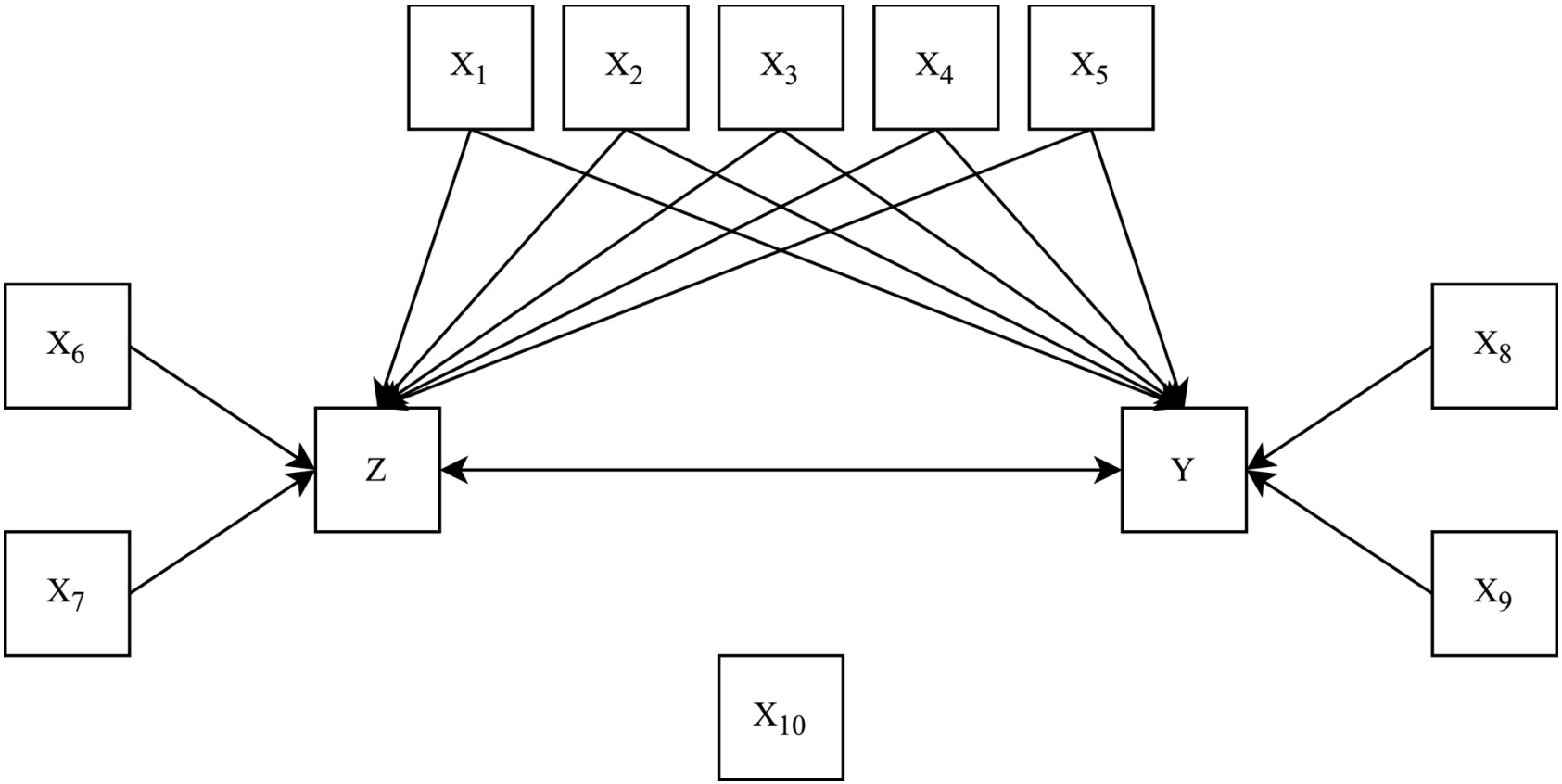

Figure 1. Diagram of Data Generation Process for the 10 Covariate Condition

Note: Z = treatment indicator, Y = outcome, $X_1 - X_5$ = true confounders, X_6 and X_7 = treatment predictors, X_8 and X_9 = outcome predictors, X_{10} = unrelated to treatment or outcome.

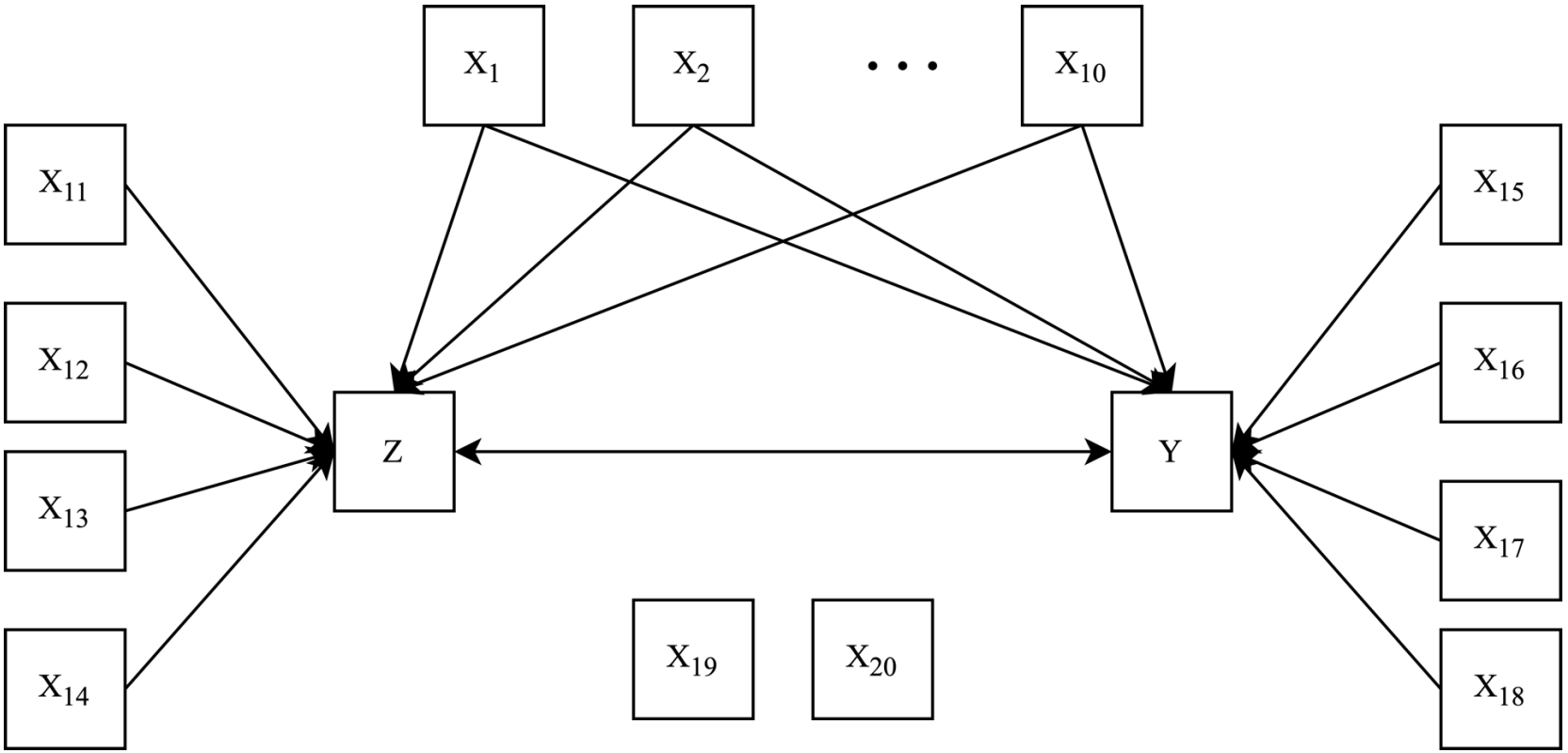

Figure 2. Diagram of Data Generation Process for the 20 Covariate Condition

Note: Z = treatment indicator, Y = outcome, $X_1 - X_{10}$ = true confounders, $X_{11} - X_{14}$ = treatment predictors, $X_{15} - X_{18}$ = outcome predictors, $X_{19} - X_{20}$ = unrelated to treatment or outcome.

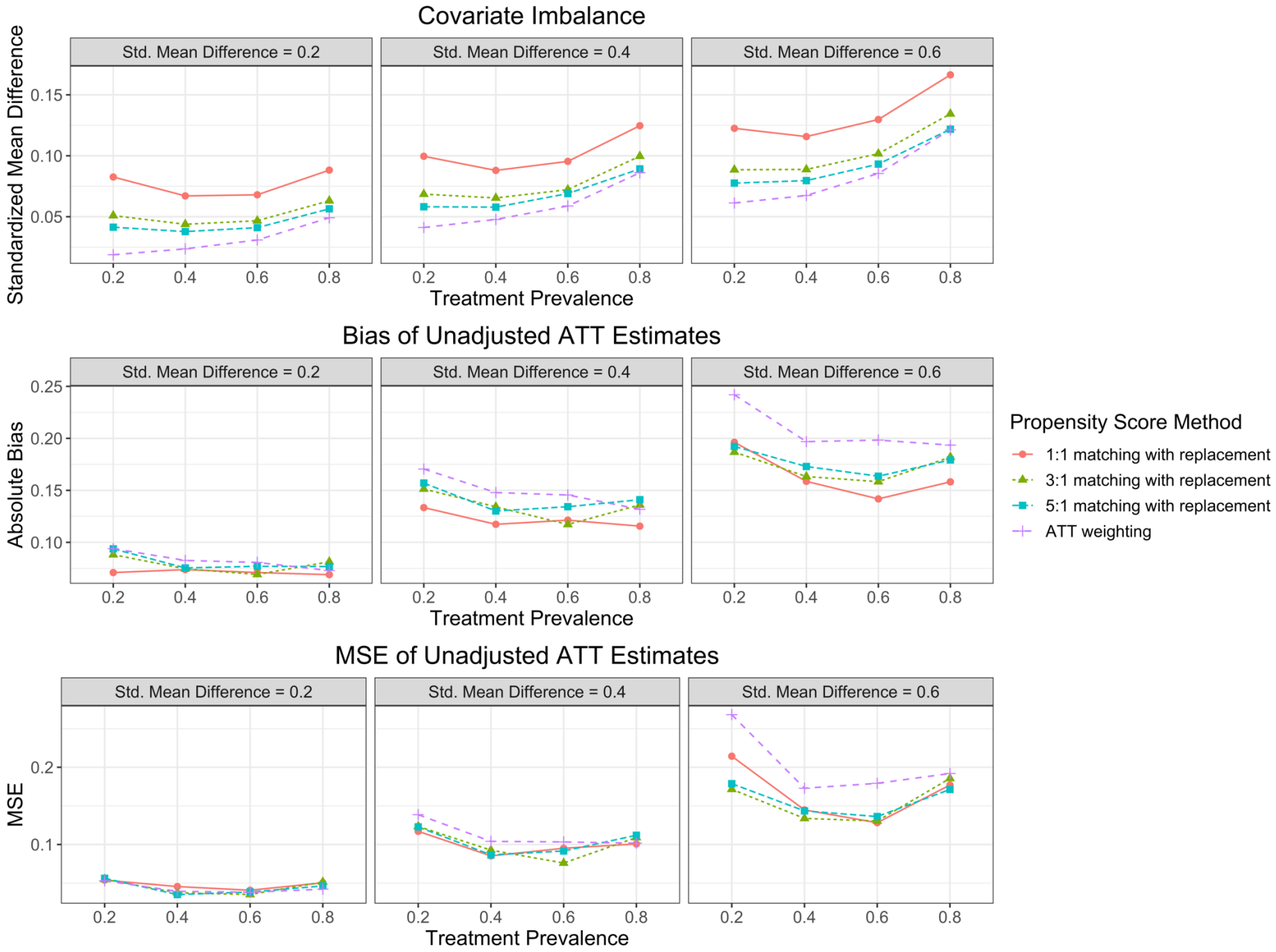

Figure 3. Covariate Imbalance, Bias, and MSE After Matching or Weighting

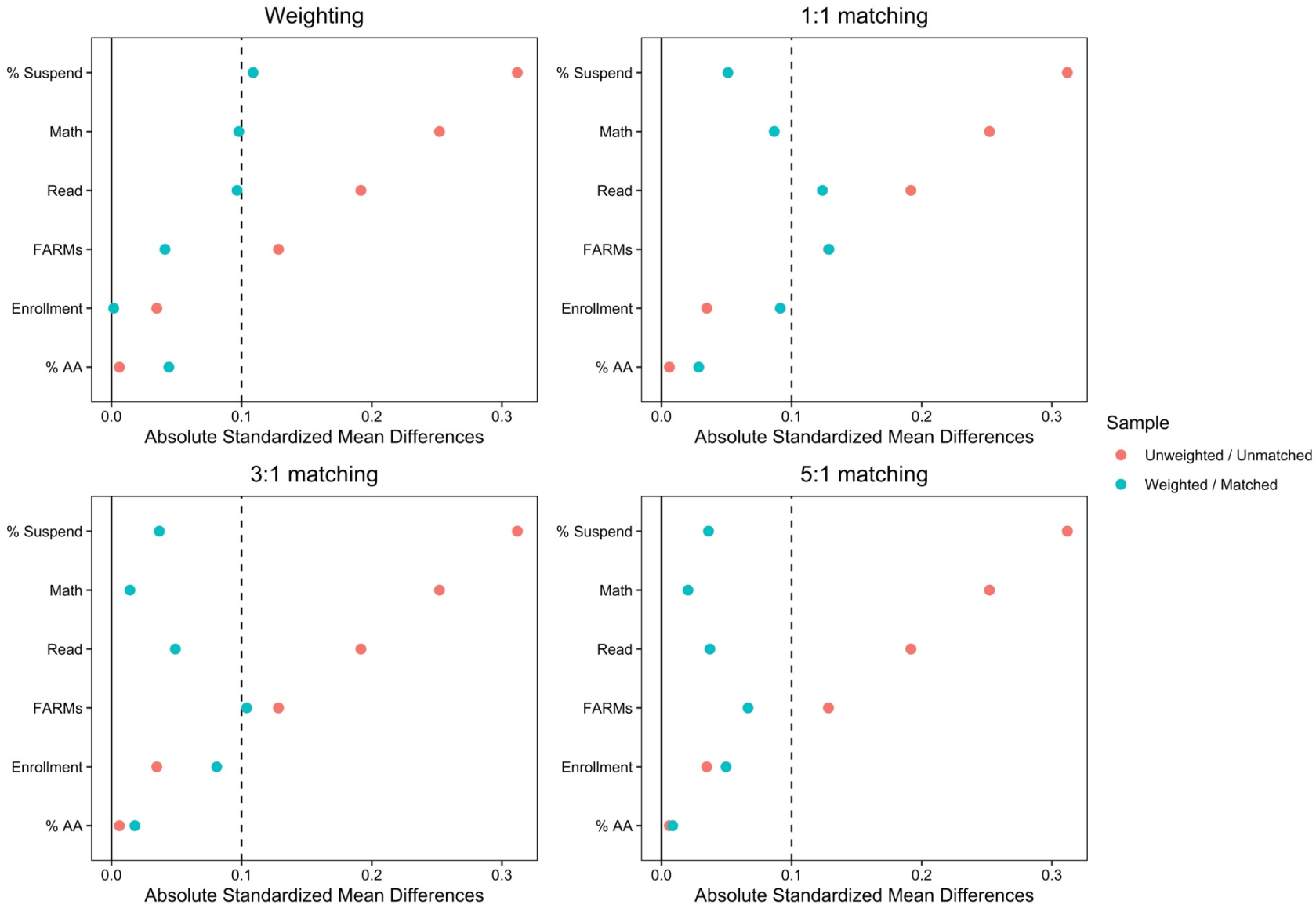

Figure 4. 2007-08 School Year Covariate Imbalance Before and After Matching or Weighting

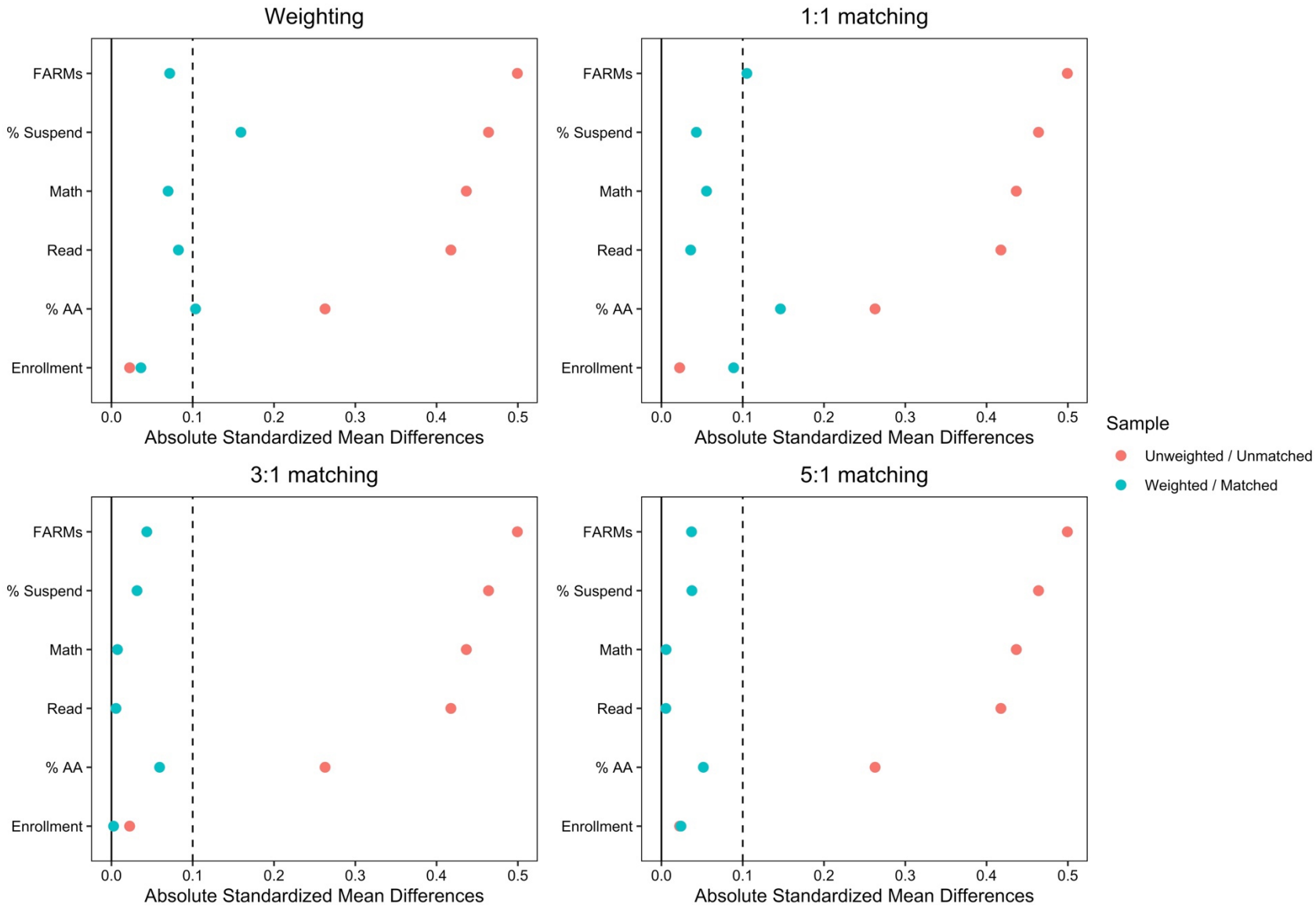

Figure 5. 2013-14 School Year Covariate Imbalance Before and After Matching or Weighting