ABSTRACT

Title of Dissertation:	THE PERFORMANCE OF BALANCE DIAGNOSTICS FOR PROPENSITY-SCORE MATCHED SAMPLES IN MULTILEVEL SETTINGS
	Alyson Burnett, Doctor of Philosophy, 2019
Dissertation directed by:	Laura M. Stapleton, Professor Department of Human Development and Quantitative Methodology, Measurement, Statistics and Evaluation Program

The purpose of the study was to assess and demonstrate the use of diagnostics for samples matched with propensity scores in multilevel settings. A Monte Carlo simulation was conducted that assessed the ability of different balance measures to identify the correctly specified propensity score model and predict bias in treatment effect estimates. The balance diagnostics included absolute standardized bias (ASB) and variance ratios calculated across the pooled sample as well as the same balance measures calculated separately for each cluster and then summarized across the sample (within-cluster balance measures). The results indicated that overall across conditions, the pooled ASB was most effective for predicting treatment effect bias but the within-cluster ASB (summarized as a median across clusters) was most effective for identifying the correctly specified model. However, many of the within-cluster balance measures were not feasible with small cluster sizes. Empirical illustrations

from two distinct datasets demonstrated the different approaches to modeling, matching, and assessing balance in a multilevel setting depending on the cluster size. The dissertation concludes with a discussion of limitations, implications, and topics for further research.

THE PERFORMANCE OF BALANCE DIAGNOSTICS FOR PROPENSITY-SCORE MATCHED SAMPLES IN MULTILEVEL SETTINGS

by

Alyson Burnett

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
2019

Advisory Committee:
Professor Laura M. Stapleton, Chair
Professor Gregory R. Hancock
Professor Jeffrey Harring
Professor Ana Taboada Barber
Professor David Blazar

Dedication

To my husband, Thomas Burnett, for believing in me and working tirelessly to support me in achieving this goal.

To Michael and Suzanne Burnett, Richard and Tammy Knapp, and Alexandra Limas, who helped us raise our son while I took on this endeavor.

To my advisor and mentor, Laura Stapleton, for her excellent feedback, encouragement, and support over the course of my graduate career at Maryland.

This would not have been possible without each of you!

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

A primary objective in evaluation research is to establish a cause-and-effect relation between a policy or program and its intended outcomes. Providing such evidence on the effectiveness of a policy or program, which will be referred to as "treatment" in this dissertation, is important for both formative and summative purposes. In education, program evaluations can help district leaders, principals, and teachers determine whether a treatment is working as intended and make informed decisions about how to adapt it to make it more effective. Evaluations can also help funders, such as the U.S. Department of Education or foundations, determine whether to continue funding a treatment or to invest in others. Over the last decade, the federal government has heavily invested in systematic reviews in fields such as education (What Works Clearinghouse, U.S. Department of Education), home visiting (Home Visiting Evidence of Effectiveness, U.S. Department of Health and Human Services), teen pregnancy prevention (Teen Pregnancy Prevention Evidence Review, U.S. Department of Health and Human Services), and labor (Clearinghouse for Labor Evaluation and Research, U.S. Department of Labor) that summarize the causal effects of a treatment, and these reviews often determine which programs receive funding.

As a result, evaluation researchers are increasingly interested in how to design studies so that they can establish a causal relation between the treatment and its intended outcomes. Research design features, such as how individuals came to receive their treatment and the similarity of individuals across treatment conditions, are the basis for whether cause can be established. Ideally, individuals are randomly assigned to treatment conditions so that any variables that might confound with the treatment are randomly

distributed across both conditions. However, this is not always feasible or desirable in applied settings in which individuals may voluntarily elect to receive a treatment or are selected based on specific criteria. An added complication to designing applied evaluation studies is that individuals are often nested together within organizational structures, such as schools, and the selection process and implementation of the treatment may vary across those schools. Such data structures typically require the use of multilevel models to account for variations in the outcome and the relation between predictor variables and the outcome across clusters.

Propensity score (PS) matching is a useful approach for designing studies with comparable treatment and control groups when randomization is not feasible. A PS is a balancing score that represents the propensity that a unit is selected for treatment (Rosenbaum & Rubin, 1983). PS matching involves a four step process: (1) modeling the PS using variables that are related to treatment selection and/or the outcome, (2) matching treated units to control units with similar PS estimates, (3) performing diagnostics on the matched sample, and (4) estimating the treatment effect with the matched sample. The diagnostic step is even more critical with PS matching than with a randomized controlled trial because the researcher must make a convincing case that the resulting treatment effect estimates are unbiased. To do so, the researcher must evaluate whether the units in the treatment and control groups have similar means and distributions of the measured covariates, a feature known as balance.

Although PS matching has recently been extended to multilevel settings, it is not yet clear how to apply diagnostic procedures to nested data structures. The study described in this dissertation tested several possible measures for evaluating balance of

multilevel PS-matched samples. The balance measures were evaluated based on two criteria: 1) ability to identify the best fitting PS model and 2) ability to predict bias in the treatment effect estimate. Findings from the study expand the literature on methods for multilevel PS matching and provide guidance to researchers wanting to assess balance under several multilevel contexts.

Using a Monte Carlo simulation, the study compares various methods of summarizing balance measures in multilevel settings. Two important balance measures in single-level settings are absolute standardized bias (ASB) and variance ratios. ASB measures the absolute standardized difference in means between treatment and control units on the covariate of interest, and a variance ratio is simply the ratio of the variance of the treatment group to the variance of the control group on the covariate. When PS matching is implemented in multilevel settings, these balance measures may be pooled across the sample while ignoring cluster membership, or they may be calculated separately within each cluster and then summarized. For example, a researcher taking the pooled approach would calculate the ASB for the full sample, whereas a researcher taking the within-cluster approach may calculate the ASB for each cluster and then report the mean or median of the cluster ASBs. Researchers undertaking PS matching in multilevel settings have used both approaches, yet neither had been previously corroborated by methodological research. The dissertation tested several variations of both pooled and within-cluster ASB and variance ratios for evaluating balance in studies that utilize multilevel PS matching.

Based on prior research, I hypothesized that the preferred balance measure should depend on several factors, including the size of the clusters, the value of the intracluster

correlation coefficients (ICCs) of the unit-level covariates, the extent of the misspecification of the PS model, and the matching method. These factors were manipulated in the Monte Carlo simulation to better understand the conditions in which different balance measures would be useful. More detailed hypotheses are provided at the end of Chapter 2.

In addition to the methodological study, I also conducted two empirical illustrations of multilevel PS matching using real data. The first illustration used student achievement data clustered at the classroom level, and the second used a worldwide youth health survey clustered at the country level. Both illustrations demonstrate the four steps in PS matching with a multilevel dataset, while using the recommended balance summary measures from the simulation study based on the cluster size and other characteristics of the data. This not only helped to verify that the recommended procedures are feasible to implement with real datasets but also identified additional challenges in applied settings that should be considered in future research. The empirical illustrations serve as models to applied researchers wishing to implement PS matching in similar multilevel contexts.

The dissertation is divided into six chapters. The next chapter lays out the conceptual framework and reviews the literature on PS matching with the recent expansion to multilevel settings. Chapter 3 describes the design of the simulation study that assesses balance measures for multilevel PS matching, and Chapter 4 provides the simulation results. Chapter 5 then shifts the focus to the empirical illustrations and includes a brief background on those datasets and a description of the empirical methods and results. Finally, the dissertation concludes in Chapter 6 with a summary of the results

from the simulation and the empirical illustrations, a discussion of the implications of the findings, acknowledgement of the study's limitations, and suggestions for future research.

Chapter 2. Review of the Literature

As described in the previous chapter, the purpose of the dissertation is to better understand diagnostic tools that can be used to assess covariate balance when implementing PS matching with multilevel data. This chapter aims to expound upon the literature motivating the study, drawing together research in both educational and medical statistics and insights from methodological and applied studies. It begins with an introduction to causal inference and the invention of PS methods for establishing cause in observational studies. It then explains each of the steps required for implementing PS methods, including modeling the PS, conditioning on the PS, performing diagnostics, and estimating the TE. The chapter then describes approaches for using PS methods in multilevel contexts, including the current gaps in this literature. Finally, the chapter concludes with research questions for the Monte Carlo simulation study to investigate the use of diagnostic measures for multilevel PS matching.

2.1 Potential Outcomes Framework

The fundamental problem of causality is that we can only observe one potential outcome for each person (Holland, 1986). The counterfactual, or the unobserved outcome, is unknown because the same person cannot simultaneously serve as both the treatment and the control. A person receiving treatment can be compared to a person not receiving the treatment, or can be compared to himself at another point in time. Rubin's (1974) model for causal inference is the one most often used in statistics and social sciences to understand this problem and how it can be resolved (Schafer & Kang, 2008). To understand this model, a few key terms must be defined. In the equations that follow, $Y_i(D_i)$ represents the potential outcomes for each individual i. In the case of a binary

treatment indicator, D_i equals 1 for person *i* in the treatment condition and 0 for person *i* in the control.

First, the individual treatment effect (ITE), is the difference between the outcomes for an individual receiving treatment compared to if he or she had not received treatment. The ITE, which cannot be determined, can be written as follows:

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

Because the ITE cannot be observed, the average treatment effect (ATE), is typically of interest. This is the expected value of the ITE over the population:

$$\tau_{ATE} = E[Y(1) - Y(0)] \tag{2}$$

However, the ATE is not always of interest, because the treatment may be designed for a very specific group of people and the effect should not be estimated for those whom the treatment was not intended. In this case, the average treatment effect on the treated (ATT) may be of greater interest because it focuses explicitly on the treatment effect (TE) for those who received treatment. It is defined as:

$$\tau_{ATT} = E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1] \tag{3}$$

As with the counterfactual of the ITE, the counterfactual of the ATT cannot be observed, since only those in the treatment group are of interest and they cannot receive two conditions at once. As such, researchers interested in the ATT must find an adequate substitute for the counterfactual that allows them to meet the assumptions outlined below (Caliendo & Kopeinig, 2008).

2.1.1 Assumptions for causal inference. In order to estimate the ATE or ATT without bias, one must meet several assumptions (Rubin, 1978; 1980). First, one must assume that the treatment is the same for all individuals and that a treatment applied to

one individual does not affect the outcome of another individual (Rubin, 1980). This set of assumptions is referred to as the stable unit treatment value assumption (SUTVA). As will be discussed in later sections of this chapter, SUTVA is unlikely to hold in multilevel studies unless the researcher accounts for this design feature in the analysis.

Second, there should be no unmeasured confounders, an assumption known as unconfoundedness (Rubin, 1978). This can also be thought of as an independence assumption, as it requires that treatment assignment and potential outcomes are independent (Holland, 1986). This assumption can be met through randomization of treatment status or through conditioning on variables. Once all covariates that could influence treatment assignment and the outcomes are incorporated into the TE model, any differences on the outcome between those who receive treatment and those who receive the counterfactual is solely due to treatment status.

Finally, one must meet the assumption of common support, also known as overlap. That is, there is a positive probability of receiving both the treatment and the control for all possible values of the covariates (Rosenbaum & Rubin, 1983). If certain individuals have a 0 probability of receiving a condition, then it is not possible to estimate their causal effects because the alternative would not be possible for them. Empirical studies typically define common support in terms of the overlap in PS distributions and discard any units with PS estimates outside the range of the opposite group (Stuart, 2010). Together, Rosenbaum and Rubin refer to the unconfoundedness and common support assumptions as "strong ignorability."

2.1.2 Importance of design in causal inference. Randomized controlled trials (RCTs) can meet the assumptions of strong ignorability and common support by nature

of the randomization. If participants are randomly assigned to the treatment condition, then everyone has a positive probability of selection for either the treatment or the control and thus meets the assumption of common support. Moreover, when participants are randomly assigned to treatment, any variables that could influence the outcome are, on expectation, randomly distributed across treatment and control groups and thus cannot be confounds. Because of these two features, the ATE/ATT estimate can be directly measured as the difference between the average outcome in the treatment group and the average outcome in the control group. In the case of an RCT, the ATE and ATT estimates are equivalent because treated individuals will not differ systematically from the overall population (Austin, 2011). Although adding covariates to the TE model can improve the precision of the estimate, they are not needed to meet the assumptions for causal inference. Observational studies, also known as quasi-experiments, have the same goal of establishing a causal relation between a treatment and an outcome, but unlike RCTs, individuals are not randomly assigned to treatment conditions (Cochran, 1965). As such, the ATE and ATT are not assumed to be equivalent, and the TE cannot be estimated through direct comparison of treatment and control participants (Austin).

When randomization is not feasible, one can make unbiased causal inferences through use of a variety of methods, for example regression discontinuity design (RDD) or an interrupted time series (ITS), a special case of an RDD. To use a discontinuity design, specific conditions must be met (Murnane & Willett, 2011). First, participants should be arrayed along an underlying continuum—called a forcing variable—that is related to the outcome of interest. Second, there should be an exogenously determined cut-point or threshold that divides participants into treatment groups. Third, there should

be a reliable and valid outcome of interest. In these studies, the analyst makes causal inferences based on whether there is a discontinuity in the relation between the forcing variable and the outcome at the cut-point. For example, Gormley, Gayer, Phillips, and Dawson (2005) used an RDD to show the effect of universal preschool on achievement measures using age as the forcing variable and the birthday cutoff as the cut-point. They could then compare the effects of attending preschool between those whose birthdays were just before the cutoff to those whose birthdays were just after the cutoff and had to wait another year before attending. In the case of an ITS design, the forcing variable is time, and the cut-point is a sudden change of policy. With an ITS, outcome data must be collected at many times before and after the cut-point in order to establish cause. For example, Wagenaar, Maldonado-Molina, and Wagenaar (2009) used an interrupted time series to analyze the effects of alcohol tax increases in Alaska on alcohol-related disease mortality from 1976 to 2004. While these methods help to infer cause, they are not appropriate in all situations.

With certain datasets and research questions, neither an RCT nor a discontinuity design, such as RDD and ITS, are feasible. For example, this would occur in an observational study in which there is not an exogenous cut-point along a forcing variable. In these circumstances, the researcher must account for the nonrandom treatment assignment and ensure that treatment assignment and potential outcomes are independent by conditioning on certain variables through use of regression-based adjustments, matching, or stratification. As one can imagine, each of these options for analyzing observational designs becomes increasingly complicated as more variables are needed in order to meet the assumption of unconfoundedness. In the case of regression-based

approaches, a model with a very large number of covariates may become over-parameterized and difficult to fit, especially if the sample size is not large. With matching, incorporating a large number of factors on which to match could lead to very few matches between treatment and control individuals. Likewise, stratification based on a large number of factors may lead to too many strata from which to estimate effects. However, methods such as matching and stratification are feasible approaches for minimizing selection bias in nonrandomized studies with the use of a single balancing score, or PS, that incorporates a large set of variables (Rosenbaum & Rubin, 1983).

2.2 Propensity Score Methods

Rosenbaum and Rubin (1983), who first introduced the PS, described it as a balancing score. The score is formed using regression, typically logistic or probit regression, with treatment status regressed on all relevant variables that are likely to predict treatment status and/or the outcome. In this paper, Rosenbaum and Rubin demonstrate that if treatment status is considered to be strongly ignorable given a set of covariates (in other words, there are no remaining confounds once the set of covariates are included), then the treatment status is also considered to be strongly ignorable given a PS that incorporates these covariates. Once the treatment status is considered to be strongly ignorable, the difference between the treatment and control means at any value of the PS is an unbiased estimate of the TE. As such, the PS can be used to produce unbiased TE estimates.

Since Rosenbaum and Rubin introduced the theory of propensity scores, PS methods have become increasingly popular in the social sciences. A recent literature review on PS methods showed nearly exponential growth in the number of articles

published using PS methods between 1991 and 2009 (Thoemmes & Kim, 2011). The largest percentage of articles were published in the field of education, but other fields included public health, criminology, psychology, sociology, social work, and family studies. In the future, PS methods will likely expand to other fields, such as business or engineering, that seek to make decisions based on the success of a practice. For example, a grocery store could use PS methods to determine the effects of self-checkout machines on customer satisfaction. In this scenario, customers who had the option of either using the self-checkout line or the traditional line may be asked to complete a survey after leaving the store. Using data from the survey, customers who used the self-checkout line could be matched with customers with similar characteristics who used the traditional line. To form the PS model, the survey would include questions to predict which line a customer chose, such as age, number of items purchased, experience using a self-checkout machine, and number of produce items without barcodes. After matching customers on these characteristics, they could then be compared to give an unbiased estimate of satisfaction between those using self-checkout and traditional lines.

There are three primary advantages to using PS methods that may be influencing their growing popularity. First, as previously mentioned, PS methods are particularly useful when a large number of covariates are needed to meet the assumption of strong ignorability (Rubin & Thomas, 1996; Shadish, Clark, & Steiner, 2008). In a review of studies that used PS methods, researchers used an average of 31 covariates in their model, but some used well over 100 covariates (Thoemmes & Kim, 2011). Matching or stratifying on each of these variables separately would prove to be nearly impossible.

With PS methods, one can more easily incorporate a large number of covariates and examine the balance of the covariates across treatment and control groups.

Second, PS methods separate the design from the analysis, and as such, they can be used to better design an observational study before the analysis stage (Austin, 2011; Shadish et al., 2008). For example, one can examine the degree of overlap between the PS estimates in the treatment and control groups to determine whether certain individuals should be removed from the sample in order to meet the assumption of common support. One can also check the balance of the covariates before and after matching or stratification to ensure that the PS model was specified correctly prior to the analysis of outcomes. Such tweaks to the PS model would be made separately from the specification of the outcome model; therefore, the researcher would not be biased to adjust the design features after reviewing the results from the outcome model. These diagnostics are also much simpler to assess when the PS model is separate from the outcome model (Austin). Goodness-of-fit measures used in OLS regression, such as the model's R^2, will not provide information on whether balance on the covariates has been achieved.

Third, empirical research demonstrates that PS methods are close to approximating an RCT when certain conditions are met. Using data from the National Supported Work demonstration, Lalonde (1986) compared the results from the RCT to those from observational study techniques using another, non-random, control group. The observational methods included covariate adjustment, difference-in-differences analysis that compares the change in earnings before and after training between treatment and control groups, and a difference-in-differences analysis that also included covariate adjustment. The difference-in-differences analysis that controlled for pre-training

differences and demographic variables yielded results similar to the experimental results; however, other techniques that did not control for all confounds were biased. Later, Dehejia and Wahba (1999) expanded on this work by comparing the experimental results to the observational study results using PS methods. They found that the TE estimates using PS methods were much closer to the experimental results than the other observational methods. The authors concluded that PS methods can be a substitute for RCTs to estimate treatment impact as long as the variables that predict treatment assignment can be measured and there is sufficient overlap in propensity scores (see discussion of overlap in 2.2.3).

When implementing a PS method, researchers must undertake a series of steps and make key decisions within each step. These can be summarized into four key steps, each with their own set of sub-steps and decisions: 1) modeling the PS, 2) implementing the selected PS method, 3) performing diagnostics, and 4) estimating the TE[1].

2.2.1 Step 1: Modeling the propensity score. When modeling the PS, the typical decision in the case of binary treatment is whether to use a logit or a probit model, both of which are designed to handle a dichotomous dependent variable by fitting a nonlinear function to the data. A logistic regression uses a logit link function, which, assuming a single predictor, can be written as:

$$F(x) = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x_1)}} \tag{4}$$

which can be rewritten as the inverse of the logistic function, *g*, as follows:

[1] Caliendo and Kopeinig (2008) also include sensitivity analyses as a fifth step, including sensitivity tests for the unconfoundedness and common support assumptions.

$$g(F(x)) = \ln(\frac{F(x)}{1-F(x)}) = \beta_0 + \beta_1 x_1 \quad (5)$$

Where F(x) is the probability that the unit received treatment given the linear combination of the predictors, base e is the exponential function, ln is the natural logarithm, β_0 is the intercept, and $\beta_1 x_1$ is the regression coefficient multiplied by a predictor.

Probit regression uses an inverse normal link function, which can be written as follows:

$$F(x) = \Phi(\beta_0 + \beta_1 x_1) \quad (6)$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution.

Research suggests that when used for the purpose of creating PS estimates, these models yield similar results, so the choice is not critical (Caliendo & Kopeinig, 2008). A systematic review of 86 studies that used PS methods showed that 78 percent used logistic regression, 12 percent used probit regression, and the rest were unclear about the type of model used (Thoemmes & Kim, 2011). However, a logistic model can be interpreted in terms of odds ratios, whereas the probit model does not have a direct interpretation. For this reason and because logistic regression is more widely used, a logit model may be more interpretable to the target audience of the research. Other non-parametric methods, such as boosted modeling have been proposed (McCaffrey, Ridgeway, & Morral, 2004), but are seldom used. Boosted modeling is a multivariate nonparametric regression technique that is more flexible than parametric regression because it does not assume that the relation between each covariate and treatment selection is linear and additive on the log-odds scale. Instead, it uses an algorithm to

automatically model a nonlinear relationship between a dependent variable (in this case, treatment status) and a large number of covariates (McCaffrey et al.). Lee, Lessler, and Stuart (2010) showed that using a boosted model outperformed logistic regression PS models in terms of bias reduction and 95% confidence interval coverage.

Despite the recent advances with boosted modeling, logistic regression is still the default modeling option for researchers implementing PS methods, including multilevel PS methods. All methodological studies that have assessed multilevel PS methods have used logistic regression models. For this reason, studying balance measures for multilevel PS matching under the assumption that a researcher has used a logistic regression model is more relevant and understandable to both the methodological and applied research communities. Furthermore, because the purpose of this research is to test balance measures, rather than modeling techniques, the use of logistic or boosted modeling is not important for answering the research questions. Either modeling approach could be used with the balance measures. Therefore, the remainder of this dissertation assumes the use of logistic regression for PS modeling.

By contrast, choosing which variables to include in the PS model is a rather important decision for ensuring that the assumption of unconfoundedness is met. If systematic differences exist between the treatment and control groups on confounders that are not included in the PS model, then TE estimates will be biased. A confounder is a variable associated with both treatment status and the outcome (Austin, Grootendorst, & Anderson, 2007). Although researchers agree on the importance of selecting appropriate variables, they differ in their guidance on how to select them and how many to select. For example, Caliendo and Kopeinig (2008) recommend including any variables that

influence *both* the treatment status and the outcome (true confounders), but Schafer and Kang (2008) recommend including any variables that influence *either* the treatment status *or* the outcome. However, both sets of authors emphasize the importance of understanding theory and previous research on the relations between variables and the outcomes and having institutional knowledge about how participants are sorted into treatment conditions when selecting variables to include. A series of simulations by Austin, Grootendorst, and Anderson (2007) compared variable balance and reduction in TE estimate bias of four approaches to selecting variables: selecting confounders only, selecting only variables associated with the outcome, selecting all measured variables, and selecting all variables associated with treatment selection. The selection techniques were equivalent in terms of achieving balanced samples, but omitting a confounder led to biased TE estimates. It is typical that researchers may only have access to common demographic variables such as age, race/ethnicity, gender, and a measure of social-economic status, but using these exclusively rather than variables guided by theory will lead to biased TE estimates because it is likely that a confounder will be missed (Thoemmes & Kim, 2011). Furthermore, researchers should not remove predictors based on statistical significance, because the purpose of the model is not to achieve parsimony, but rather, to achieve balance between treatment and control groups (Schafer & Kang).

Once the appropriate variables are selected, the researcher would then need to choose the functional forms of the variables, for example whether to include any polynomial or interaction terms. However, research suggests that once the confounders are included, slight deviations of the PS model will have minimal impacts on selection bias (Drake, 1993; Waernbaum, 2010). Drake (1993) conducted a series of simulations

that varied the misspecifications of the PS model and the outcome model, which was estimated through stratification. The true PS model was a quadratic logistic model and misspecifications included a linear logistic model and omitting a quadratic term. Drake found that misspecifications of the outcome model led to much greater biases in the TE than did misspecifications of the PS model. Similarly, Waernbaum found in a series of simulations that misspecifications of the PS, such as omitting higher order terms, did not increase TE estimate bias in a matching design. As will be discussed in subsequent sections of this chapter, modeling decisions are more critical when implementing PS methods with multilevel data.

2.2.2 Step 2: Implementing the propensity score method. Once the PS estimates have been obtained using the logistic or probit regression functions (equations 4-6), researchers can use them in one of four types of PS methods: 1) matching, 2) stratification, 3) inverse probability of treatment weighting (IPTW), or 4) covariate adjustment (Austin, 2011). This step is often referred to as conditioning on the PS (Austin et al., 2007). In matching, treatment and control units are matched that have the same or the most similar PS estimates, and the matched sample can then be used to estimate the ATT (Imbens, 2004). In stratification, the sample is ordered based on PS estimates and then subdivided into a number of equal-sized strata, either based on the total number of individuals in the sample (to estimate the ATE) or the total number of treated individuals (to estimate the ATT; Imbens). The TE is then estimated for each stratum and then averaged to calculate an overall TE. Applying IPTW is similar to applying sampling weights. In estimating the ATE, each unit's weight is equal to the inverse probability of receiving the treatment that they actually received (weights can be modified for

estimating the ATT if it is of interest). Finally, using PS estimates for covariate adjustment simply means that instead of using a large number of separate covariates in an outcome model, the researcher would instead use the PS estimate as a single covariate in the outcome model.

In introducing the theory of PS methods, Rosenbaum and Rubin (1983) argue that matching, stratification, and covariate adjustments using PS estimates can produce unbiased estimates (they did not consider the use of inverse probability weights in their paper). However, there may be advantages to choosing one method over another. For example, matching, stratification, and inverse probability weighting have the advantage of separating the design from the analysis, allowing one to directly estimate the TE once the PS model has been specified (Austin, 2011). Although the PS estimates are formed as part of a separate step from the outcome model in the case of covariate adjustment, the researcher must still fit a regression model that predicts the outcome based on the PS estimate and treatment status. In doing so, there might be temptations to adjust the model to make the expected outcome more likely. Research also suggests that some methods are preferable to others in terms of achieving precise TE estimates, achieving balance across covariates, and removing bias in the TE estimates. In comparing the precision and bias of TE estimates, Schafer and Kang (2008) found that PS stratification and PS covariate adjustment were more effective than using inverse probability weights for measures of the ATE. Another series of simulation studies found that PS matching led to greater covariate balance between treatment and control units than did stratification, presumably for estimation of the ATT (Austin et al., 2007). However, there is also much variation in the effectiveness within each method, depending on how it is implemented. For example,

with stratification, the researcher must decide how many strata to use, and this has implications on the precision and bias of the TE estimates. The remainder of this chapter focuses on matching, since this method is used in the majority of applied studies that utilize PS methods (Thoemmes & Kim, 2011). Furthermore, nearly all methodological studies that investigated multilevel PS methods focused on matching.

When implementing PS matching, researchers must make four decisions regarding the matching algorithm: (1) whether to match with or without replacement, (2) whether to match 1 to 1 or many to 1, (3) whether to use a caliper, and (4) whether to use nearest neighbor or another matching estimator such as optimal matching. The most intuitive approach is 1:1 nearest neighbor matching without replacement in which treatment units are matched to the nearest control unit. Once matched, the control units are no longer available for other matches, and unmatched control units are discarded. Because the quality of the matches may change based on the order in which units are matched, it is recommended that matches are made in a random order (Caliendo & Kopeinig, 2008). Several adaptations can be made to the simple 1:1 nearest neighbor matching approach to either improve matches or limit the reduction in sample size. Researchers may decide to sample with replacement rather than sampling without replacement. This means that once a control unit has been matched with a treatment unit, the same control unit can be matched with another treatment unit if it is the nearest neighbor. This can improve the overall quality of the matches, but decreases precision of the TE estimate because there are fewer distinct individuals included in the sample (Caliendo & Kopeinig). Similarly, researchers may decide to match multiple control units to the same treatment unit (*k*:1 nearest neighbor matching), a decision that has the same

tradeoffs between bias and precision. As more control units are matched to the same treatment unit, the quality of each match decreases but the precision of the TE estimate increases because there are more individuals included in the sample. Researchers using this approach must decide how many control units should be allowed to match to each treatment unit. For either matching with replacement or *k*:1 matching, weights should be applied in the outcome analysis to account for individuals being in the sample more than once or for oversampling. Another adaptation to nearest neighbor matching is to limit matches to those that are within a specified distance from the treatment unit. This means that some treatment units that do not have control units within the specified distance, or caliper, would not be matched or included in the analysis. Many researchers use a caliper of .2 standard deviations of the PS, because it has been shown to be effective for removing selection bias (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1985). As expected, applying a caliper can improve the quality of matches but can also increase variance and decrease power by removing individuals from the sample.

Another form of matching that a researcher may choose is optimal matching. Rather than focusing on the best match for an individual treatment unit as in nearest neighbor matching, it instead considers the overall quality of all matches (Stuart, 2010). In optimal matching, each match is chosen to minimize a measure of global distance. A simulation study that compared nearest neighbor matching to optimal matching showed that the two approaches performed similarly in terms of achieving covariate balance across treatment groups and minimizing propensity distances between matched pairs (Gu & Rosenbaum, 1993). Nearest neighbor matching may be preferred in some contexts and

fields because it can be more easily explained to an audience unfamiliar with PS matching techniques.

To summarize step 2, researchers wishing to use PS estimates may choose from one of four methods: matching, stratification, inverse probability weights, and covariate adjustment. Once the method is selected, more decisions are required. Matching is the most common method and is thus the focus of this review. For matching, one must decide on the particular matching algorithm, including whether to match with or without replacement, to match one to one or one to many, whether to use a caliper, and whether to use nearest neighbor matching or another matching estimator.

2.2.3 Step 3: Performing diagnostics. Once the PS method has been implemented, researchers must examine two properties 1) the balance property and 2) the region of common support (Thoemmes & Kim, 2011). The balance property assesses—either numerically or graphically—whether the treatment and control groups have similar sample means and distributions on the covariates. This section first describes the numeric summaries and then describes the graphical displays that can be used to evaluate covariate balance.

There are several possible numeric diagnostics for evaluating balance; such methods include calculating standardized mean differences between treatment and control groups before and after matching, conducting t-tests to compare treatment and control groups after matching or within strata, examining the ratio of the variances of the PS estimates in the treatment and control groups, examining the ratio of the variances of the residuals orthogonal to the PS estimates in the treatment and control groups for each covariate, and comparing the pseudo-R^2 before and after matching (Caliendo &

Kopeinig, 2008; Stuart, 2010). Although the most popular method is the t-test approach (Thoemmes & Kim, 2011), Stuart warns that this is problematic for two reasons. First, while balance is a within-sample characteristic, hypothesis tests refer to a broader population from which the sample was drawn. Second, the change from a significant difference before matching to a non-significant difference after matching could be due to a loss in power due to trimming the sample, rather than to an improvement in balance.

Calculating the standardized mean difference, also referred to as standardized bias or Cohen's *d* (Cohen, 1988), is another popular choice for evaluating balance in PS matched or stratified samples. The standardized mean difference can be calculated as

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}} \tag{7}$$

where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ are the sample means on the covariate of the treatment and control groups, respectively, and $s^2_{treatment}$ and $s^2_{control}$ are their respective variances. A slight variation of this formula is to use the variance among treatment group members exclusively rather than the pooled variance (Stuart, 2010); however, there is no consensus in the literature about which variance is more appropriate for evaluating balance. In the case of PS matching, one should calculate the standardized mean difference for each covariate before and after matching using the same variance for both (Stuart, 2010). A benefit of using the standardized mean difference for evaluating balance is that it can be evaluated against a predetermined threshold. However, one must consult the literature in the particular field of study to select an appropriate threshold; recommendations for thresholds may be as conservative as .05 (U.S. Department of Education, 2017) or as liberal as .25 (Harder, Stuart, & Anthony, 2010) depending on the field and the purpose

of the research. Ho, Imai, King, and Stuart (2007) argue that the level of acceptable standardized bias should depend on the importance of the covariate in predicting the treatment assignment and outcome measure, where higher levels of bias are acceptable for covariates of lower importance.

Austin (2009) showed in a set of simulations that balance measures that evaluate PS matching should incorporate the distribution of the covariates rather than just means, as with standardized mean differences. This can be achieved through measuring the ratio of variances between treatment and control groups as follows:

$$ratio = \frac{s^2_{treatment}}{s^2_{control}} \tag{8}$$

Ratios close to 1 indicate greater balance between treated and untreated subjects on the covariate. In Austin's (2009) simulations, the ratio of variances outperformed the standardized mean differences for detecting bias in the TE estimate. The standardized mean differences for the correctly specified PS model and a misspecified model that did not include a confounder included in the model were both small, indicating little bias, but the ratio of variances were further from 1 with the misspecified model in comparison to the correctly specified model. As with standardized mean differences, the ratio of variances can be evaluated against set criteria. For example, Rubin (2001) considered a ratio of variance below .5 or above 2 as too extreme. However, although quantitative methodologists recommend examining the ratio of variances, it is not yet a common practice in applied research (Thoemmes & Kim, 2011).

Graphics for evaluating balance of the covariates include quantile-quantile (QQ) plots and a plot of the standardized mean differences before and after matching (if matching is used). QQ plots compare the quantiles of a variable for the treatment group

on one axis and the corresponding quantile for the control group on the opposite axis. When the distributions are balanced, the dots will track along the 45 degree line. Figure 1 provides an example of QQ plots for three variables before and after matching. All three variables show improved balance after matching, as the dots track more closely to the 45 degree line. The middle plots show that two units have been matched even though they have different values on the dichotomous variable, which may or may not be acceptable to the researcher depending on the importance of the variable for predicting the treatment assignment and outcome.

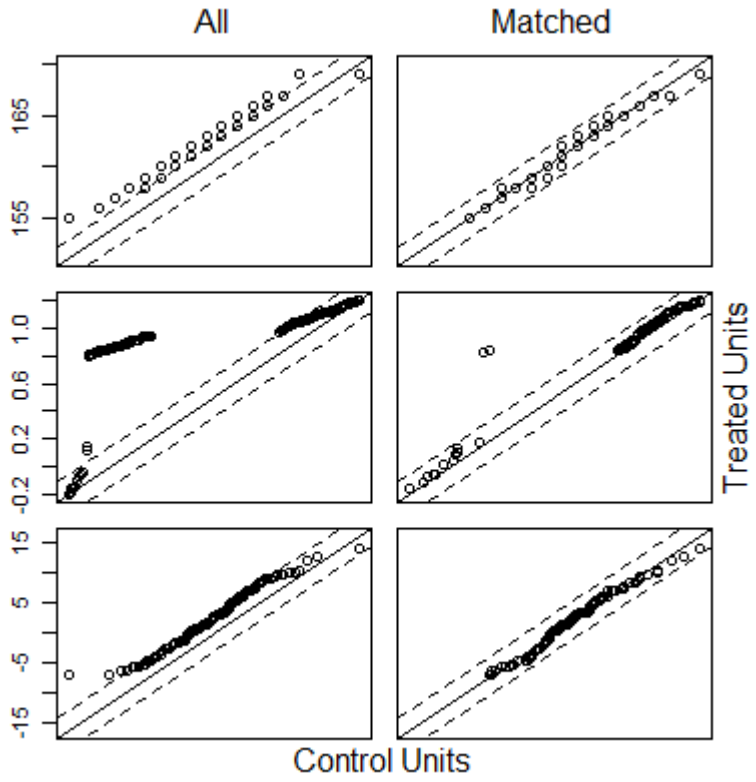

Figure 1. QQ plots before and after matching for three variables

The standardized mean difference plot shows all covariates together, which allows one to visually examine the degree to which bias was reduced for each covariate. It may be the case that although matching reduced bias overall, it increased for certain variables, so researchers can use such a plot to identify those variables and determine

whether such an increase is tolerable (Stuart, 2010). In this example (Figure 2), the standardized mean difference decreased for all variables except one, which was determined to be an acceptable level of balance for that particular variable, because it was not believed to be strongly related to the outcome.

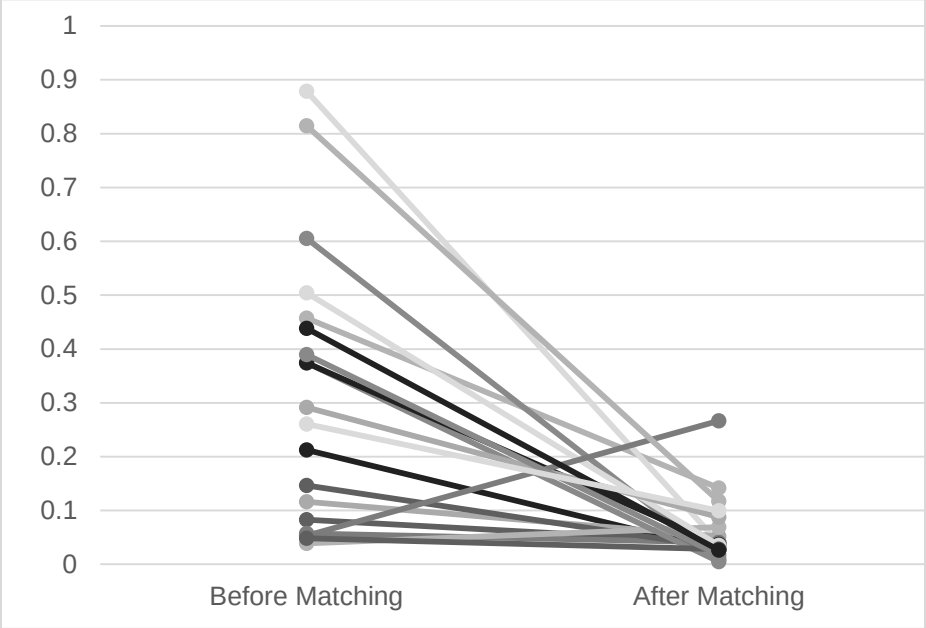

Figure 2. Absolute standardized mean differences of covariates before and after matching with a caliper width of .25

Ho et al. (2007) explain that assessing balance is an iterative process. Researchers should not just choose one matching method and assess it for balance to confirm their approach and then move on. Instead, they should compare the balance of several variations of matching or stratification methods (e.g., optimal or nearest neighbor, one-to-one or one-to-many) and models (e.g., including higher order terms or interactions) and then select the combination that achieves the greatest level of balance. During this iterative process, one should not select models based on statistical significance of the estimated regression coefficients, because the primary objective is to achieve balanced samples (Austin, 2011).

In sum, balance can be assessed numerically, ideally through standardized mean differences and variance ratios, and graphically, through use of QQ plots and standardized mean difference plots. These procedures should be carried out in an iterative process to select the methods and models that will achieve the greatest level of balance and thus minimize bias.

To ensure that treatment and control groups are comparable for estimating the ATT and ATE, one must also evaluate common support through examining the region of overlap in the distributions of PS estimates. In the case of PS matching, common support is typically assessed through use of a "jitter plot" that illustrates the distribution of PS estimates for all matched and unmatched units. This plot is divided into four categories: unmatched treatment units (if any), matched treatment units, matched control units, and unmatched control units. Ideally, any treatment or control units that are much higher or lower than units in the opposite group should not be matched, because this would indicate a lack of common support. Figure 1 illustrates a sample with an acceptable level of common support through the use of 1:1 nearest neighbor matching with a caliper of .2 standard deviations. In this case, the caliper rule effectively removed the majority of the control units and a few treatment units because there were not enough comparable units in the opposite group.

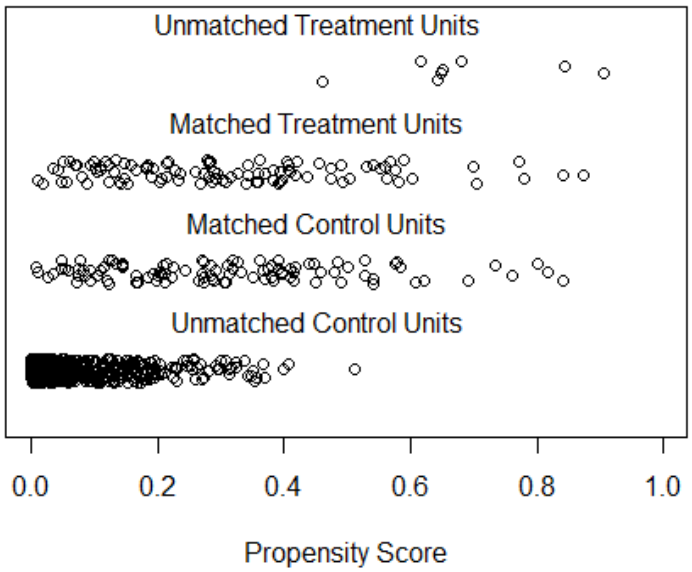

Figure 3. Jitter plot used to examine evidence for common support

Similar plots may be used to evaluate common support in PS stratification and weighting studies. In the case of stratification, one may construct a plot with the treatment and control units on separate horizontal lines as in Figure 1 but with vertical lines that divide the plot into strata to evaluate overlap within each stratum. For weighting, one may consider creating a plot in which the dot size represents the weight of the unit in the analysis. Numeric diagnostics for overlap include the simple comparison of the PS minima and maxima across matched treatment and control units and estimation of the region of overlap using nonparametric kernel densities (Smith & Todd, 2005).

Researchers may address the problem of lack of common support in several ways. If using PS matching, they can improve the level of common support by applying a caliper or narrowing the caliper width, as described in the above example of Figure 1. Another approach that can be used with any type of PS method is to apply a trimming rule. For example, one may remove any individuals with PS estimates smaller than the minima or larger than the maxima of the opposite group (if the ATE is of interest), or only remove control group members who are below the minimum or above the maximum

of the treatment group (if the ATT is of interest; Caliendo & Kopeinig, 2008). Crump, Hotz, Imbens, and Mitnik (2009) proposed a trimming rule that removes all units with PS estimates below .1 or above .9; however, they warn that applying a trimming rule may decrease the external validity by focusing on a smaller subset of the originally identified population. It may also change the estimand of interest. If many control or treatment units need to be discarded because there are no nearby units in the opposite group, then it may not be possible to estimate the ATE. Likewise, the researcher may not be able to estimate the ATT if treated units need to be discarded because there are no nearby control units (Stuart, 2010). In these cases, the researcher may need to select a different dataset to answer the particular research questions because the groups are too different to produce unbiased TE estimates (Rubin, 2001).

As will be discussed later in this chapter, it is not yet clear how researchers should apply PS diagnostics in multilevel studies, such as when students are nested within schools. Researchers would need to know whether to perform diagnostics for each school separately or to perform tests that pool all of the schools together. For example, one could calculate separate standardized mean differences for each covariate within each school, or one could calculate standardized mean differences for each covariate, aggregating across schools. No current literature has clarified how these different approaches would impact detecting bias and making adjustments to the PS modeling or conditioning approach.

2.2.4 Step 4: Estimating the treatment effect. After the propensity model has been selected based on the results of diagnostics, the final step is to use the PS estimates in the TE model. With matching methods, one can calculate the average outcome in each group with the matched sample using weights as needed to account for matching with

replacement or matching to multiple control units. In the MatchIt package of R, all unmatched units have a weight of 0 and matched treatment units have a weight of 1 (Ho, Imai, King, & Stuart, 2017). The control weights are calculated in three steps. First, thinking of matching in terms of creating groups with at least one treated unit and at least one control unit, a preliminary weight is calculated by dividing the number of treated units by the number of control units in the group. Second, if the same control unit was used across multiple groups, then the weights are summed across them. Third, the control group weights are rescaled such that the sum of all of the weights equals the number of uniquely matched pairs (Ho et al.). Using the weights, the ATT can be estimated as:

$$ATT = \frac{\sum_{i=1}^{n} w_{it} y_{it}}{\sum_{i=1}^{n} w_{it}} - \frac{\sum_{i=1}^{n} w_{ic} y_{ic}}{\sum_{i=1}^{n} w_{ic}} \tag{9}$$

where w_{it} and w_{ic} are the weights and y_{it} and y_{ic} are the values on the response variable for group *i* in the treatment and control groups, respectively. In the case of stratification, the treatment effect of each stratum is first estimated and then aggregated across strata. Weights should be applied based on the size of each stratum, and these weights will determine the type of treatment effect estimate. If the ATT is of interest, weights should be based on the number of treatment units in the stratum, but if the ATE is of interest, they should be based on the total number of treated and untreated units in the stratum, as follows:

$$ATE = \sum_{i=1}^{n} w_i (y_{it} - y_{ic}) \tag{10}$$

where w_i is the weight assigned to stratum *i*, and y_{it} and y_{ic} are the values on the response variable for stratum *i* in the treatment and control groups, respectively

The most contested topic regarding TE estimation with designs that utilize PS matching is whether to use variance estimates that account for the matched nature of the data (Stuart, 2010). Matched pairs will likely be correlated on the outcome measures, but the research is unclear on whether PS matched samples should be treated as dependent samples for the TE analysis. While some researchers argue that it is not necessary (e.g., Schafer & Kang, 2008), others have shown through simulations that accounting for the matched nature of the data in the variance estimates leads to more precise estimates of the TE (e.g., Gayat, Resche-Rigon, Mary, & Porcher, 2012). One way of accounting for matching in the variance estimates is through bootstrap methods, which are used to estimate the sampling variability of parameters (Austin & Small, 2014). Given that this issue is still contested in the literature, I opted to ignore the dependencies for the purpose of this study, given that the focus is on balance measures during the diagnostic stage and incorporating the bootstrap methods during the TE estimation stage would be unlikely to affect their performance.

Another issue when considering TE analysis is whether to include any covariates that are already being accounted for in the propensity scores. As previously discussed, Rosenbaum and Rubin (1983) showed that as long as the PS incorporates all confounds, the difference between the treatment and control means at any value of the PS is an unbiased estimate of the TE. This means that the TE analysis does not need to include covariates if the PS model is correctly specified. However, since it is impossible to know whether the model is correctly specified, incorporating covariates into the TE analysis may be beneficial. Incorporating covariates into both the PS model and the TE model is known as doubly robust estimation (Robins, Rotnitzky, & Zhao, 1994). Funk et al. (2011)

showed that when doubly robust estimation is applied, only one of the two models needs to be correctly specified to obtain unbiased treatment effect estimates.

Including covariates in the TE model may have additional benefits. First, including covariates in the TE model will explain a greater proportion of the total variance of the outcome, which will increase the power for detecting a significant effect. Second, covariates are useful for understanding how the treatment interacts with other variables, for example, the effects of a reading intervention may vary according to baseline reading ability. Third, in some cases, PS matching reduces the balance of some variables even though it improves balance overall. Incorporating these variables into the TE analysis would provide greater assurance that the TE estimates are unbiased. **2.2.5 Summary of the four steps.** Implementing a PS method entails following a series of steps and making specific decisions within each step. First, the researcher must model the PS, which involves determining whether to use a logit or probit model, the variables to include, and the functional forms of those variables. Next, the researcher should select a PS method—either matching, stratification, inverse probability weighting, or regression adjustment—and determine the particular algorithm for the method, for example choosing nearest neighbor or optimal matching. Third, the researcher should assess balance and overlap and iterate with different PS models and conditioning approaches until an approach is selected that will minimize TE estimate bias. In the final step, the researcher uses the PS estimates in the outcome model and must determine the weights and variance estimates to apply based on the particular PS approach. The next section will discuss the expansion of PS methods into multilevel settings and review the literature on how the four steps are applied to various multilevel contexts.

2.3 Multilevel Propensity Score Matching

Although PS matching has gained popularity as a way to make causal inferences in observational studies, researchers are only beginning to use them in multilevel settings, such as when students are nested within schools, and have used a wide variety of approaches (Arpino & Cannas, 2016). A series of empirical studies by Hong and colleagues on the effects of kindergarten retention on academic and social outcomes illustrate that there is no one best approach for all multilevel studies using PS methods (Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007, 2008). These studies used stratification, but the modeling and conditioning approaches could be applied to multilevel PS matching studies as well. Across these multilevel studies, the authors employed different PS methods depending on the research questions at hand. For example, to answer questions about whether a school's retention policy had an effect on children on average at the school, the authors stratified the schools in the sample based on a PS model that predicted the probability of a school allowing retention to estimate the ATE (Hong & Raudenbush, 2005). They did not include student-level characteristics in the PS model or stratify at the student level because the question was about the effect of a school-level policy on school-level outcomes. Other studies that investigated the ATE for students across schools used multilevel models to estimate the propensity of being retained based on individual, classroom, and/or school characteristics (Hong & Raudenbush, 2006; Hong & Yu, 2007, 2008). For example, one study examined the effect of being retained in schools with low retention rates separately from the effect of being retained in schools with high retention rates (Hong & Raudenbush, 2006). To do so, the authors first divided schools into low and high retention schools and within each

school type, they formed a separate multilevel PS model that incorporated school and student-level characteristics. They then used the PS estimates to divide students into strata and to estimate the ATE separately for low retention and high retention schools using multilevel regression models. The authors explained that without randomization of the school-level retention rate, the propensity of retention under a low-retention rate for children attending high-retention schools and the propensity of retention under a high-retention rate for children attending low-retention schools were not estimable. Another study investigated the effects of retention for students with a risk of being retained (Hong & Yu, 2007). The study utilized a three-level PS model that predicted retention based on student, classroom, and school-level characteristics. Children who had 0 probability of being retained were removed from the sample, and the remaining were pooled together across schools and stratified based on the PS estimate for the TE analysis. The reading and math outcomes were estimating using a three-level model. Although these studies all explored the effects of retention on kindergarten outcomes, the specific research questions warranted different approaches to dealing with the nested nature of the data.

As demonstrated in the Hong studies, a researcher may employ a variety of PS modeling and conditioning approaches depending on the level of treatment assignment and the research questions of interest. When treatment is assigned to clusters, as in the first example (Hong & Raudenbush, 2005), the propensity score should reflect the probability of the cluster being assigned to treatment. This means that the researcher will select the cluster-level variables that are likely to predict treatment assignment and the outcome of interest to include in the PS model. Unit-level variables would not need to be

included because they do not predict treatment status, and therefore, a single-level PS model at the cluster level with matching between clusters is sufficient.

Research questions about unit-level treatment within multilevel contexts are more complex, and may fall into these general categories:

1) What is the overall TE across clusters?

2) What is the cluster-level TE on average and does it vary across clusters?

3) What unit-level factors moderate the TE?

4) What cluster-level factors moderate the TE (cross-level interaction)?

Depending on the research question, specific models will be required. For example, as will be discussed later in the chapter, the need to estimate a cluster-specific TE has implications for both the type of PS model that should be used and whether matching should be conducted within or across clusters (e.g., Thoemmes & West, 2011; Rickles & Seltzer, 2014; Arpino & Mealli, 2011; Kim & Seltzer, 2007). The next section of this chapter will focus on the decisions that must be made when using PS matching in multilevel settings. The current research has focused solely on the decisions related to modeling the PS (step 1) and matching units using the PS estimates (step 2, implementing the PS method). These decisions go hand-in-hand such that one must consider the matching approach while choosing the most appropriate PS model, and likewise should consider the PS model while choosing the most appropriate matching approach. Although these studies have utilized various balance diagnostics (step 3) to assess the modeling and matching approaches tested, none have specifically studied the use of balance diagnostics with multilevel PS matching. More research is needed to determine the performance of various assessments of covariate balance for detecting bias in TE estimates. Potential

approaches for assessing covariate balance will be discussed following the review of the literature on modeling and matching approaches for studies utilizing multilevel PS matching.

Research suggests that there are four primary types of models that can be used for PS estimation in multilevel settings when treatment status is at the individual level: 1) a single individual-level (SL) model that ignores clustering, 2) a fixed effects (FE) model, 3) a multilevel model with random intercepts only (RI model), and 4) a multilevel model with random intercepts and slopes (RIS model; Thoemmes & West, 2011). These models may be paired with three types of matching approaches when implementing a PS method: 1) pooled matching, 2) within-cluster matching, or two-stage matching (Rickles & Seltzer, 2014), which is a hybrid of the first two approaches. The paragraphs that follow will describe each modeling and matching approach and then will discuss the research on the optimal combinations in various settings.

2.3.1 Propensity score models for multilevel settings (step 1). The simplest PS model is an SL model that does not include any cluster-level covariates or account for any differences in the selection process across clusters (see Equations 4-6). The SL model ignores the presence of clustering; however, using such a model does not mean that clustering is not accounted for in the PS method. For example, a researcher may use an SL model to estimate the PS and then account for the clustered data by matching units within clusters and/or using a multilevel model in the TE estimation.

Multilevel PS models take into account cluster-level differences in treatment assignment (e.g., policies that affect the likelihood of being retained). The main consideration with using a multilevel PS model is whether to only allow the intercepts

(the probability of being selected for treatment) to vary across clusters by using the RI model or whether to allow both the intercepts and the slopes (the relation between the covariates and the treatment assignment status) to vary by using the RIS model. The RI model is represented in equation 11 below, and the RIS model is represented in equation 12 (Thoemmes & West, 2011).

$$\text{logit}(\text{P}_{ij}) = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{ij} + \sum_{q=1}^{Q} \gamma_{0q} W_j + \sum_{i=1}^{I} \gamma_{1i} W_j X_{ij} + u_{0j} \tag{11}$$

$$\text{logit}(\text{P}_{ij}) = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{ij} + \sum_{q=1}^{Q} \gamma_{0q} W_j + \sum_{i=1}^{I} \gamma_{1i} W_j X_{ij} + u_{0j} + \sum_{p=1}^{P} X_{ij} u_{1j} \tag{12}$$

In equation 11, $\text{logit}(\text{P}_{ij})$ represents the estimated logit of the PS for the *i*th unit in the *j*th cluster, γ_{00} represents an intercept, γ_{p0} represents the regression coefficients for the individual-level covariates, γ_{0q} represents the regression coefficients for the cluster-level covariates, γ_{1i} represents the regression coefficients of any interactions between individual-level and cluster-level covariates, and u_{0j} is the random effects component influencing the intercept of each cluster, *j*. In Equation 12, u_{1j} represents the cluster-level random effect components influencing the regression slopes of the individual-level covariates.

A slight variation of the RI model is an SL model that incorporates each cluster as a separate predictor using dummy variables, which Thoemmes and West (2011) refer to as a fixed effects (FE) model.

$$\text{logit}(\text{P}_i) = \beta_0 + \sum_{p=1}^{P} \beta_p X_{ip} + \sum_{q=1}^{Q} \beta_q C_{iq} + \sum_{i=1}^{I} \beta_i C_{iq} X_{ip} \tag{13}$$

In this equation, β_0 represents an intercept, β_p represents regression coefficients for the individual-level covariates, β_q represents regression coefficients for the dummy coded variables indicating cluster membership, and β_i represents regression coefficients for potential interactions between the clusters and the individual-level predictors.

This option is preferred to using the same SL model across all clusters, because it accounts for cluster-level differences in the selection process. As indicated by the equation, the model may include interactions between clusters and individual-level variables, but this option is only available in limited situations because it requires large sample sizes within each cluster (Thoemmes & West, 2011). Although Thoemmes and West do not specify the cluster size needed for the model, there would need to at least be more individuals in a cluster than variables so that the researcher trusts the cluster-specific regression coefficients. In cases in which there are no cluster-level predictors or interactions between clusters and individual-level predictors, the fit of the RI model and the fixed effects model would be equivalent (Kim & Seltzer, 2007).

2.3.2 Matching with propensity scores in multilevel settings (step 2). As described previously, once the PS model is specified, researchers then need to implement a PS method, typically either matching or stratification. In multilevel contexts, in addition to the other decisions that need to be made during this stage, researchers must also decide whether to condition the PS estimates within or across clusters. For example, in the case of matching, the researcher could decide to restrict a student's match to only other students in the same school or to allow the student to match to students in other schools. Three matching approaches have been examined in simulation studies: (1) pooled

matching, (2) within-cluster matching, and (3) two-stage matching (Rickles & Seltzer, 2014).

The within-cluster matching approach was first demonstrated by Rosenbaum (1986) in a study on the effect of dropping out of high school. The advantage of this approach is that cluster-level covariates do not need to be included in the PS model, because they will be the same for each student within the same school. This helps to ensure that the unconfoundedness assumption is met for cluster-level confounders, because it is not possible to leave out any important cluster-level covariates (Rickles & Seltzer, 2014). However, the disadvantage is that there may not be a close match within the same school if there is a small number of treatment students, control students, or both (Kelcey, 2011). Restricting matches to within the same school would potentially lead to either bad matches resulting in biased TE estimates, or, if using a caliper, would significantly reduce sample size and therefore power to detect a significant TE. Another concern is that if not all of the treatment students can be matched because of the restriction to matches within their schools, the estimand changes because the ATT estimate does not actually represent the full population of treatment individuals and no longer reflects the relative cluster sizes (Arpino & Cannas, 2016).

In many situations, matching within the same cluster is not feasible because the clusters are too small or because there are no comparable treatment and control units within the same cluster. For example, in a study of the effect of being retained in kindergarten, there may not be any students with similar characteristics as the students who were retained in the same school. As such, these studies of kindergarten retention tend to pool together kindergarteners across schools and implement PS stratification

based on a multilevel PS model with both student- and school-level predictors (e.g., Hong & Yu, 2008). However, when matching occurs across clusters, cluster-level differences in how units come to receive their treatment are not accounted for unless they are explicitly included in the PS model. As such, studies that utilize pooled matching should incorporate a multilevel PS model with a comprehensive list of cluster-level covariates (Hong & Yu). Another drawback to pooled matching is that one cannot preserve the design of a multi-site study where treatment and control students are compared within the same cluster (Rickles & Seltzer, 2014). Furthermore, the researcher cannot estimate a within-cluster TE and/or heterogeneity of the TE if that is of interest.

In the field of educational statistics, Stuart and Rubin (2008) proposed a two-stage matching approach for treatments implemented in just one school, and Rickles and Seltzer (2014) extended the approach for treatments implemented across many schools. The purpose of this approach is to preserve the conceptual design of a multisite study while circumventing the problem of small sample sizes. In this approach, treatment students are matched to control students within the same school if there is an adequate match within the same caliper. If there are no adequate matches within the same school, then the treatment student is matched to a control student from another school that has similar school-level characteristics. Rickles and Seltzer describe the two-stage approach as occurring in three steps: matching, adjustment, and analysis. First, the authors match students, ideally to control students within the same school, but to control students in similar schools if needed. Second, for matches made outside of the school, the authors make an adjustment to estimate what the outcome would have been for the student if they were in the same school as the treatment student. Third, they estimate the ATT within

each school, across schools, and review TE variation. One could take many different approaches to identifying similar schools, but in Rickles and Seltzer's empirical example, they divided schools into quantiles on an achievement index, which was based on schools' average standardized test scores. Students could then only be matched with students from schools within the same quintile. Separately, in the field of biostatistics, Aprino and Cannas (2016) developed a similar method, which they call "preferential within-cluster matching" in which they first attempt to match a unit within the same cluster and then move to the pooled dataset if a unit within the specified caliper width is not available in the unit's cluster. However, unlike Rickles and Seltzer, Arpino and Cannas do not restrict matches to clusters with similar characteristics or implement an adjustment for matches made outside of the treated unit's cluster.

2.3.3 Comparison of modeling and matching approaches. Researchers have several options for accounting for multilevel data when implementing PS methods. Though not recommended in most circumstances, they may choose to ignore the multilevel structure by pairing an SL model with pooled matching. If they wish to take the multilevel data structure into account, they can do so in either the modeling stage, matching stage, or in both stages. For example, a researcher may consider pairing a multilevel PS model with pooled matching or an SL, PS model with within-cluster matching. Using Monte Carlo simulations and examples from applied datasets, researchers have compared various combinations of modeling and matching approaches for multilevel data in terms of bias of the TE estimate, covariate balance as measured by standardized bias, root mean squared error or mean squared error, and proportion of matched units. The results of these studies show that the optimal combination of

modeling and matching approach depends on several factors including the extent of variation in the treatment selection process across clusters, within cluster sample sizes, and whether balance is desired across the sample as a whole or within clusters.

Arpino and Cannas (2016) refer to the SL model with pooled matching as the "naïve approach" because it ignores clustering in both stages. Simulations and empirical examples that compare this approach to other approaches demonstrate that ignoring clustering leads to poorer outcomes in almost all circumstances (Arpino & Cannas; Arpino & Mealli, 2011; Li, Zaslavsky, & Landrum, 2013; Thoemmes & West, 2011). For example, Arpino and Mealli conducted a series of Monte Carlo simulations that compared an SL propensity score model to two types of multilevel models that were each paired with pooled matching. The data generating model included three individual-level covariates and one cluster-level covariate, which was omitted from the PS models. SL propensity score models underperformed multilevel models in terms of covariate balance, bias of the TE estimates, and mean squared error. The extent of improvement using a multilevel model depended on the correlation between the omitted cluster-level variable and other variables. When the omitted cluster-level variable was highly correlated with treatment status, the bias of the SL model increased in comparison to the multilevel models. Li et al. conducted a simulation with a similar design to Arpino and Maelli but used IPTW rather than matching for estimating the ATE. The study tested three PS models in combination with three alternative formulas for calculating the IPTW with clustered data. They found that ignoring the clustering of the data in the PS model and the IPTW calculation led to larger bias and root mean squared error (RMSE); however, ignoring the clustering in the IPTW calculation was more detrimental than ignoring the

clustering of the data in the PS model. Thoemmes and West conducted a simulation that tested four types of PS models (SL, fixed effects, RI, and RIS) crossed with two conditioning approaches (pooled and within clusters), two different ICCs of the individual-level covariates (.05 and .5), and two different sample sizes. The authors considered both stratification and matching in their study, but the simulation only tested stratification. For pooled stratification, they formed 10 strata across the whole sample, ignoring clusters, and for within-cluster stratification, they formed 10 strata within each cluster. Although in most cases the naïve approach led to higher levels of bias of the TE estimate and mean squared error, it was not the case in the low ICC conditions. When the ICCs of the individual-level covariates were close to 0, all modeling and conditioning approaches performed similarly.

Methodological researchers have considered whether it is best to account for variations in treatment selection in the modeling or matching stage by comparing the implementation of a multilevel PS model with pooled matching with an SL propensity score model with matching within clusters (Arpino & Cannas, 2016; Rickles & Seltzer, 2014; Thoemmes & West, 2011). The optimal approach depends on the cluster size. As previously described, if cluster sizes are small, then fewer treatment units can typically be matched when using within-cluster matching, and this changes the estimand of the ATT (Arpino & Cannas). Arpino and Cannas also explain that within-cluster matching with an SL propensity score model can be seriously biased when cluster sizes are small because of the reduction in matched units. For example, Thoemmes and West observed in an applied dataset that matching within clusters resulted in sample sizes that were on average only 5% of the original sample size and thus resulted in very large variations in

TE estimates. In simulations conducted by Arpino and Cannas, bias in the TE estimate based on within-cluster matching with an SL propensity score model started to be acceptable when the clusters had at least 300 units. When clusters were large (300 units or larger), within-cluster matching with an SL propensity score model performed better than a multilevel PS model with pooled matching, but all methods that took clustering into account were superior to the naïve approach. By contrast, when clusters had an average of 50 units each, the RI model with pooled matching and the SL model with within-cluster matching had higher levels of TE estimate bias than even the naïve approach; in this case, the fixed effects model was preferred. Rickles and Seltzer found that within-cluster matching paired with an SL propensity score model had low levels of bias in the TE estimate with slightly smaller cluster sizes. In their simulations, the cluster size was normally distributed with a mean of 200 and variance of 100.

Although Arpino and Cannas (2016) only considered the possibility of either accounting for clustered data in the PS modeling or the matching stage, other authors have considered accounting for clustering in both stages. In Thoemmes and West's (2011) simulations, using an SL instead of a multilevel PS model when matching within clusters led to a high level of bias when the ICCs of the individual-level covariates in the PS model were large, indicating large differences in the treatment selection process across clusters. Kim and Seltzer (2007) tested three types of PS models with within-cluster matching using an applied dataset and also concluded that using a multilevel model was important to reduce bias in the TE estimate. They explained that an SL propensity score model fails to achieve balance within clusters and therefore threatens the internal validity of the TE and its variation across clusters. However, these results were

inconsistent with those obtained by Rickles and Seltzer (2014) in a series of simulations that tested three different PS models with three different matching approaches. In these simulations, within cluster matching resulted in minimal levels of TE estimate bias across the different PS models (SL, RI, RIS). The authors explain that this is because within-cluster matching accounts for both observable and unobservable differences in treatment selection across clusters.

When matching within clusters is not feasible due to small cluster sizes, research suggests that two-stage matching, also known as preferential within cluster matching, outperforms pooled matching (Arpino & Cannas, 2016; Rickles & Seltzer, 2014). Rickles and Seltzer determined that the two-stage matching method proved to be the optimal method when within-cluster matching resulted in poor matches or removed too many individuals from the sample due to the caliper size. However, the performance of the two-stage method depended on pairing it with a PS model that accounted for the clustered data structure. Compared to within-cluster matching, two-stage matching led to greater bias of the TE estimate when paired with an SL propensity score model, but it performed similarly to within-cluster matching when paired with either an RI or RIS propensity score model. They also showed in an empirical example that the within-cluster matching approach removed 57% of treatment units, so the two-stage approach had better generalizability. Arpino and Cannas obtained similar results with two-stage matching, demonstrating that it performed better than within-cluster matching when clusters were small. However, they also made a distinction based on the strength of the relation between an omitted cluster-level confounder and the treatment status, which ranged from 0 to .6 in simulation conditions. When the omitted confounder had a low or medium

strength relation with treatment status, two-stage matching was preferred but when it was high, within-cluster matching was preferred.

Whether implementing within-cluster, two-stage, or pooled matching, researchers implementing a multilevel PS model need to determine whether to implement an RI or an RIS model. Simulation studies that have considered the RIS model typically only considered models with two or three individual-level covariates, all with random slopes (Rickles & Seltzer, 2014; Thoemmes & West, 2011). However, Thoemmes and West suggested that in applied studies with many more covariates, researchers should initially run the PS model with all random slopes but then remove any that are not significant. Kim and Seltzer (2007) took this approach in their applied multilevel propensity score analysis; of the 18 individual-level covariates, six had significant random slopes and were allowed to be random in the PS model. These studies suggest that both the RI model and the RIS model can work well with PS matching depending on the study design.

Kim and Seltzer (2007) explained that there are clear differences between RI and RIS settings. In RI settings, the average probability of selection differs across clusters; in RIS settings, the average probability of selection differs across clusters and the magnitudes of the slopes of multiple unit-level covariates that predict the probability of selection differs across clusters. Using real data from the Early Academic Outreach Program, Kim and Seltzer compared the balance achieved with each PS model when paired with within-cluster nearest neighbor matching with a caliper. Both models had 18 individual-level covariates, but in the RIS model, six of them were random. The results indicated that ignoring the random slopes in a RIS setting led to poorer balance within

clusters, leading to potentially biased cluster TE estimates and overestimation of between-cluster TE variation (Kim & Seltzer).

Thoemmes and West (2011) made a similar distinction between the settings in which RI models versus RIS models should be applied, referring to them as broad and narrow inference spaces—essentially the same concept as RI and RIS settings, respectively. In broad inference spaces, clustering is an incidental feature of the design, as policies for treatment assignment and delivery of the treatment are the same across clusters. As such, random slopes are not needed, and the PS analysis attempts to approximate a single-level randomized experiment in which units happen to be clustered within clusters. For example, in a federal college loan program that has the same eligibility criteria for all students in the United States, delivery is likely to be the same across all clusters and clustering is therefore incidental. By contrast, in narrow inference spaces, clustering is a central feature of the design, as different clusters have different policies of how to assign units to treatment and control conditions. In the narrow space, the PS analysis attempts to approximate a multisite randomized controlled trial and uses both random intercepts and random slopes for all of the covariates. In Thoemmes and West's simulations, the RIS propensity score model performed well in both broad and narrow inference spaces, but the RI model performed well in broad but not narrow inference spaces. The simulations operationalized broad inference spaces by setting the ICCs of the individual-level covariates to .05 and narrow inference spaces by setting the ICCs to .5. When ICCs were .05, RI and RIS models performed similarly, but when ICCs were .5, the RIS model outperformed the others across all measured outcomes (Thoemmes & West). A limitation of this simulation and others in the field is that the

study only included three individual-level covariates, which is unrealistic in empirical datasets. It would not be realistic to include random slopes for every covariate in a PS model with a large number of predictors. Another limitation, which the authors noted, is that ICCs of .5 are unrealistically high, even in narrow inference spaces.

Arpino and Maelli (2011) and Arpino and Cannas (2016) showed that the cluster sizes should also factor into the type of multilevel PS model one chooses. Both studies compared the fixed effects model to the RI model in the presence of pooled matching and an unobserved cluster-level confounder. These simulations demonstrated that the fixed effects model achieved greater balance than the RI model when the cluster sizes were small (20 or fewer units per cluster). Arpino and Cannas's simulation varied the relation between a cluster-level confounder and treatment status (0, .2, .4, and .6), which in turn caused the ICCs of treatment status to vary across conditions from .01 to .09. When clusters were small, the RI model had higher levels of imbalance and bias in the TE estimate when there was a strong relation between the omitted cluster-level confounder and treatment selection (Arpino & Cannas). By contrast, the fixed effects PS model performed reasonably well across all simulation conditions.

Several recent methodological studies provide guidance on the circumstances in which each combination of modeling and matching strategy is likely to minimize selection bias. In general, when cluster sizes are large enough to support it, using an RIS model with within-cluster matching will best reduce bias, but when cluster sizes are smaller, using two-stage matching is a good compromise between pooled and within-cluster matching. An SL model can be warranted when the treatment selection process does not vary across clusters, and an FE or RI model can be warranted when the strength

of the predictors for treatment does not vary. Although this research suggests modeling and matching choices when implementing PS methods with multilevel data, it does not yet suggest the diagnostics to perform to assess balance. The next section discusses potential approaches for assessing balance of multilevel data.

2.3.4 Balance assessment for matching with multi-level propensity scores (step 3). Each of the methodological studies on multilevel PS matching described above calculated a form of standardized bias to evaluate covariate balance but differed in whether they took a pooled or within-cluster approach to doing so (Arpino & Cannas, 2016; Arpino & Maelli, 2011; Kim & Seltzer, 2007; Rickles & Seltzer, 2014; Thoemmes & West, 2011). The pooled and within-cluster approaches to assessing balance are comparable to the pooled and within-cluster approaches for matching. In the pooled approach, clustering is ignored and standardized bias is calculated in the same way that it would be calculated in an SL study. Arpino and Maelli (2011) and Arpino and Cannas (2016) both took this approach to evaluating covariate balance by calculating the average absolute standardized bias (ASB) across clusters for each unit-level and cluster-level covariate over the Monte Carlo replications. Arpino and Maelli defined the pooled ASB as follows:

$$ASB = \left| 100 \frac{(\bar{X}_T - \bar{X}_C)}{\sqrt{.5(s_T^2 + s_C^2)}} \right| \quad (14)$$

where $\bar{X}_T$ and $\bar{X}_C$ are the sample means on the covariate of the treatment and control groups, respectively, and s_T^2 and s_C^2 are sample variances of the two groups. Arpino and Cannas (2016) used a slight variation of this formula by standardizing the difference in means with the treatment variance rather than the pooled variance.

By contrast, both Kim and Seltzer (2007) and Rickles and Seltzer (2014) took the within-cluster approach, calculating balance statistics separately for each unit-level covariate within each cluster. However, the studies took different approaches to summarizing the information. While Kim and Seltzer reported the mean differences between treatment and control units on the covariates separately for each school, Rickles and Seltzer calculated the grand-mean ASB for each covariate, as follows:

$$\frac{1}{J}\sum_{J=1}^{J} d_j \tag{15}$$

where J is the number of clusters and d_j is the ASB of the j^{th} cluster. Like the pooled ASB, the grand-mean ASB provides a single summary statistic for each covariate, but unlike the pooled ASB, the grand-mean ASB gives each cluster equal weight regardless of its size. The pooled ASB and grand-mean ASB are equivalent when all clusters have the same number of units.

Another strategy is to calculate both pooled and within-cluster balance statistics, which Thoemmes and West (2011) did in their study. For both within-cluster and pooled balance, they calculated the standardized differences on the means of each covariate and reported the median of the standardized differences. For within-cluster balance, they reported the average of the median standardized bias across all clusters and strata, and for pooled balance, they reported the median standardized bias for the unit-level and cluster-level covariates separately, averaged across all strata. For the applied example, they took the average across all covariates rather than the median. The authors did not explain why they used the median standardized bias in the simulation but the mean standardized bias in the applied example.

A limitation of each of these studies is that they did not justify their particular approaches for evaluating covariate balance although each has its strengths and weaknesses. One advantage of the pooled approach is that it provides a single summary statistic for each covariate that can be compared against a predetermined threshold. However, it may not provide enough detail if the researcher desires to achieve within-cluster balance, which is needed for reporting separate TE estimates for each cluster or reporting on cluster heterogeneity.

The within-cluster approach provides more detail for those needing to achieve balance within clusters, but may be cumbersome to review and evaluate when there are a large number of covariates and/or clusters. The researcher would then need to summarize the within-cluster balance statistics in a way that they can detect potential problems with the PS method, for example reporting the mean or median of the ASB of the covariates within each cluster, taking the grand mean of the within-cluster ASBs, or reporting the percentage of ASBs above a given threshold. As with within-cluster matching, within-cluster balance assessment is likely to only be a viable option with large within-cluster sample sizes, since estimates of standardized bias are less reliable with small samples. The incidence rate for receiving treatment within clusters also needs to be large enough to support within-cluster balance measures. Even if the cluster has 200 units, if only 2 of them receive treatment and are matched, within-cluster balance will not be a meaningful or reliable measure. For example, based on the ECLS K 2011 cohort, typically only one or two children within a school are retained in kindergarten. When the retained children are matched using 1:1 nearest neighbor matching, the within-cluster sample size is

reduced to between one and five. With this dataset and analysis, within-cluster balance cannot be estimated for many of the schools and for other schools it is not informative.

Another limitation of each of the methodological studies on multilevel PS methods is that they only used a form of standardized bias to evaluate covariate balance even though research suggests that understanding the distribution of covariates is just as important as the means (Austin, 2009). As with standardized bias, pooled or within-cluster variance ratios could be calculated in a multilevel setting. More research is needed to determine the optimal methods for calculating and summarizing covariate balance information in multilevel PS studies. A discussion of potential measures of balance in multilevel settings and ways to evaluate those measures is provided in the statement of the problem section.

2.3.5 Treatment effect estimation with multilevel propensity scores (step 4). As with single-level PS matching, the final step in multilevel PS matching is to use the matched sample in the TE estimate. The decision about what type of TE model to use when implementing PS matching is the same as with any multilevel study. The researcher would either account for the clustered nature of the data using a fixed or random effects multilevel model (see Equations 10-12) with the outcome variable regressed on treatment assignment, or would use an ordinary least squares (OLS) model with adjusted standard errors (Thoemmes & West, 2011). Treatment assignment and any other covariates would either be fixed or vary across clusters based on theoretical or empirical reasons (Thoemmes & West).

2.4 Statement of the Problem

Because the use of multilevel PS methods is still in the early stages, many questions remain. Simulation studies have helped clarify appropriate PS modeling and matching approaches under different types of multilevel contexts, and they have also clarified how these modeling and matching decisions impact covariate balance (Arpino & Cannas, 2016; Arpino & Mealli, 2011; Kim & Seltzer, 2007; Rickles & Seltzer, 2014; Thoemmes & West, 2011). However, each of these studies defined covariate balance differently. Some used a pooled approach for assessing balance (Arpino & Cannas; Arpino & Mealli), whereas others used a within-cluster approach (Kim & Seltzer; Rickles & Seltzer). (Note that Thoemmes and West reported both pooled and within-cluster balance statistics.) Moreover, the studies that took the within-cluster approach summarized the balance statistics in different ways. For example, one study provided a table with the standardized bias listed separately for each cluster, while another reported the mean standardized bias across all clusters (Kim & Seltzer; Rickles & Seltzer). More research is needed to understand which approaches for evaluating covariate balance can predict TE estimate bias in different multilevel contexts.

The question of how to assess balance in multilevel settings is more relevant to the narrow inference space in which clustering is a central feature of the study design. In narrow inference spaces, selection probabilities and characteristics that predict selection vary across clusters (Thoemmes & West, 2011), which means that some clusters may have satisfactory levels of covariate balance while others do not. If enough clusters exhibit poor levels of balance, this could lead to greater bias in the TE estimate. Furthermore, researchers studying narrow inference spaces may wish to report TE

heterogeneity or report a separate TE for each cluster, which would require the use of diagnostics at the cluster level to ensure that these estimates are not biased. By contrast, in broad inference spaces clustering is incidental to the design, and treatment selection probabilities and characteristics that predict treatment selection do not vary across clusters (Thoemmes & West). In such contexts, there is no need to estimate TE heterogeneity or the TE of individual clusters, so cluster-level balance statistics are also unnecessary. Therefore, future research on assessing covariate balance for studies that use multilevel PS matching is particularly needed for narrow inference spaces and should investigate which balance summary statistics are useful for predicting TE estimate bias.

One could imagine two types of balance measures in a multilevel setting: standardized bias and the ratio of variances of baseline covariates (Equations 7 and 8, respectively). Standardized bias is the most common metric for assessing balance in applied studies that use PS models (Thoemmes & Kim, 2011), so understanding its use in multilevel settings will be useful to applied researchers in the social sciences. Furthermore, research suggests that to detect TE estimate bias one must examine the balance of the variance of the covariates as well as the balance of the means (Austin, 2009; Rubin, 2001). Standardized bias and variance ratios are particularly useful metrics of mean and variance balance because they can be compared against pre-established thresholds.

In a multilevel PS matching study, standardized bias and variance ratios may be pooled across clusters, or they may be calculated separately for each cluster and then summarized. In pooled balance statistics, standardized mean differences and variance ratios are calculated for each covariate in the PS model of a matched sample, ignoring

cluster membership. For example, Arpino and Maelli (2011) calculated the ASB for each covariate, and their equation (14) shows that cluster membership is not factored into the calculation. To produce cluster-based summaries, standardized bias and variance ratios are first calculated for each covariate within each cluster before and after matching. With balance statistics for each covariate within each cluster, a researcher will likely need a way to summarize the information to efficiently review and act on it. For example, Rickles and Seltzer (2014) took the grand mean of the cluster-level ASB statistics (equation 15). Another strategy is to summarize across covariates or clusters using the median (Thoemmes & West, 2011), which is less sensitive to outliers compared to the mean. As a measure of the magnitude of outlying clusters, one could calculate the percentage of clusters with balance measures above commonly accepted thresholds, for example the percentage of clusters with a variance ratio below .5 or above 2. It is not yet clear whether any of these summaries of the within-cluster balance measures would be preferred to the pooled balance measures for detecting and reducing bias in TE estimates. Given the lack of investigation of balance assessment in multilevel PS applications, this study sought answers to the following questions:

1. Which pooled and within-cluster measures of variance ratios and ASB are best for selecting the correctly specified PS model? Does this vary according to ICC of the unit-level covariates, cluster size, or matching method?
2. Which pooled and within-cluster measures of variance ratios and ASB are most related to bias in the TE estimate? Does this vary according to ICC of the unit-level covariates, cluster size, PS model, or matching method?

I answered the questions using a Monte Carlo simulation and then demonstrated the use of balance measures for multilevel settings with two empirical datasets.

Several general hypotheses can be made through extrapolating from prior research on balance diagnostics in SL settings. First, I hypothesized that in most conditions variance ratio measures would perform better than ASB measures based on the results of Austin (2009), which showed through simulation that the variance ratio was more effective than ASB for detecting PS model misspecifications. Second, when comparing the different summary measures of ASB and variance ratios, I expected that the mean would perform better than the median or threshold indicators, since Stuart et al. (2013) found that the mean ASB was more strongly correlated with TE estimate bias compared to other ASB summary measures in SL settings. A similar study on PS balance by Belitser et al. (2011) suggest that these results should be moderated by sample size. Specifically, when the sample sizes are small, there was a stronger correlation between mean-based balance measures and bias compared to other balance measures, but when sample sizes were large, all tested balance measures performed similarly. Therefore, I would expect this finding to be true with smaller cluster sizes in a multilevel setting.

Other hypotheses can be made based on findings from methodological studies on multilevel PS methods. First, these studies inform us on the conditions for which to expect greater bias. For example, I expected that in the narrow inference space when the true PS model is an RIS model, using an SL model that ignores clustering would lead to biased TE estimates (Arpino & Cannas, 2016). As differences between the data generating PS model and the model imposed on the data increase, the bias in the TE estimate should also increase. Furthermore, I expected greater bias with pooled than with

within-cluster matching when cluster sizes are large, but expected greater bias with within-cluster than with pooled matching when cluster sizes were small, based on the research of Arpino and Cannas and that of Rickles and Seltzer (2014).

Understanding the conditions for which to expect greater levels of bias in the TE estimate also suggests the conditions for which balance measures can be more informative. Correlations between the balance measure and bias of the TE estimate are more meaningful with higher levels of bias. For this reason, I expected that several of the balance measures would perform similarly when bias in the TE estimate is low but that any optimal balance measures would stand apart from the others when bias in the TE estimate is higher. Multilevel PS studies also provide insight on how the ICCs of the unit-level covariates may impact the preference for pooled or within-cluster balance measures. When the ICCs are smaller, the selection process and balance should be similar across clusters, even in the presence of model misspecifications (Thoemmes & West, 2011). In this context, pooled balance measures should perform just as well as within-cluster balance measures. By contrast, when ICCs are larger, the selection process varies across clusters, causing more clusters to be imbalanced when there are misspecifications in the PS model. In this context, within-cluster balance measures may be more useful. The next chapter lays out the methods used for the simulation study in order to answer the research questions and test these particular hypotheses.

Chapter 3. Simulation Method

A simulation study was conducted to assess the ability of various balance measures to identify misspecifications in the PS model and thus potential bias in the TE estimate. This chapter describes the method used to answer the research questions outlined in Chapter 2. The chapter focuses exclusively on the simulation methods used to address the research questions; the methods used for the empirical illustrations are described in Chapter 5. The chapter begins by describing the data generation process, which involved generating data from a multilevel PS model and then generating data from the multilevel TE model. The models both included student-level covariates, a cluster-level covariate, and random intercepts and slopes. The next section describes the manipulated and fixed factors of the simulation design. The manipulated factors may be described as between-cell factors, which require running separate replications of the data, and within-cell factors, which require performing different procedures within each replicated dataset. Between cells, the simulation varied the ICCs of the student-level covariates (but not the ICC of the outcome itself) and the cluster sizes, and within cells, it varied the PS models imposed on the data, matching methods, balance measures, and the method of summarizing the balance measures across covariates. The balance measures included both pooled and within-cluster versions of ASBs and variance ratios. The chapter concludes by describing the measurement of the two outcomes: 1) use of balance measures for identifying the correctly specified PS model, and 2) correlation between the balance measures and bias in the TE estimates.

3.1 Data generation

The Monte Carlo simulation required two data generation models—the PS model, which generated the probability of being treated, and the TE model, which generated the value on the response variable, or outcome. The motivating context of the simulation is the narrow inference space in which clustering is a central feature of the study design. In the narrow inference space, the rates of treatment selection and the strength of the relation between the predictors and treatment assignment and the outcome vary across clusters. In such contexts, the RIS model (Equation 12) is appropriate for both the PS and outcome model and was therefore used for data generation.

The specific variables and parameters used in the equation are based on an empirical analysis of the effect of kindergarten retention on first grade reading outcomes using the ECLS-K: 2011 data (NCES, Tourangeau et al., 2015). This empirical illustration is one of the two illustrations described in greater detail in Chapter 5. The empirical analysis included 36 variables that were expected to be predictive of kindergarten retention and later reading achievement based on prior research on kindergarten retention. However, in order to more efficiently manipulate factors for the simulation, only the parameters for the three student-level covariates (kindergarten reading achievement, kindergarten math achievement, and age at kindergarten entry) and the one school-level covariate (number of students retained in the prior school year) that were most predictive of kindergarten retention were included in the simulation study. The choice to include a small set of variables in a simulation is a common approach in methodological studies of multilevel propensity score methods (e.g, Arpino & Maelli, 2011; Rickles & Seltzer, 2014; Thoemmes & West, 2011). Although the number of

covariates in the PS model can affect covariate balance and the level of TE estimate bias, it should not affect the relative performance of different types of balance measures. Because the goal of this study is to understand the performance of various balance measures in the context of multilevel PS matching, the choice to limit the number of covariates should make the results parsimonious and interpretable without sacrificing accuracy. However, the empirical illustrations covered in Chapter 5 provide an example of how to apply the balance measures examined in the simulation to real datasets when a larger set of covariates are included in the PS model.

The propensity scores were generated with the following model:

$$\begin{aligned}
&\text{logit}(\mathrm{T_{ij}}) = \beta_{0j} + \beta_{Rj} X_{Rij} + \beta_{Mj} X_{Mij} + \beta_{Aj} X_{Aij} \\
&\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\
&\beta_{Rj} = \gamma_{R0} + u_{Rj} \\
&\beta_{Mj} = \gamma_{M0} + u_{Mj} \\
&\beta_{Aj} = \gamma_{A0}
\end{aligned} \tag{18}$$

In the top line, β_{0j} is the intercept for the j^{th} cluster and β_{Rj}, β_{Mj}, and β_{Aj} are coefficients for their respective unit-level variables, X_R, X_M, and X_A, which represent the kindergarten reading score, kindergarten math score, and age at kindergarten entry, respectively. The next line shows that the cluster intercept is composed of γ_{00}, the grand intercept, γ_{01}, a coefficient for a school-level variable, W_j, the number of children retained from kindergarten in the prior school year, and u_{0j}, the school-level deviation from the expected value, based on the grand intercept and W_j. In the remaining lines, u_{Rj} and u_{Mj} are school-level deviations from the grand regression coefficients (γ_{R0} and γ_{M0}), indicating that the relation between each of the student-level predictors and $\mathrm{T_{ij}}$ (treatment

status represented by a student being retained in kindergarten) and between kindergarten math and T_{ij} varies randomly across schools. Because the empirical dataset indicated that there was little variation in the relation between age at kindergarten entry and kindergarten retention across schools, the intercept varies but the slopes are fixed.

The parameter values for this data generation model are as follows: $\gamma_{00} = -1$; $\gamma_{01} = .38$; $\gamma_{R0} = -1.4$; $\gamma_{M0} = -1.7$; and $\gamma_{A0} = -.16$. The parameters are based off of an RIS model performed on the ECLS-K 2011 dataset with the exception of the grand intercept (γ_{00}). In the empirical dataset, γ_{00} = -6.76, indicating that a student who attended a school with an average retention rate and who was at his or her school average on reading, math, and age at kindergarten entry would have an odds of 1:862 of being retained. This ratio would not be practical for the purpose of the simulation, since matching would need to occur within schools in some of the conditions. Therefore, for the purpose of the simulation, the parameter was changed to -1, making the odds of being retained under the same conditions to 1:2.7. Information about the distribution of the covariates and of the random effects in the PS model and the TE model are described later in the chapter.

Outcome values, scores on first grade reading, were generated based on the following multilevel linear regression model:

$$
\begin{aligned}
&Y_{ij} = \beta_{0j} + \beta_{Tj}T_{ij} + \beta_{Rj}X_{Rij} + \beta_{Mj}X_{Mij} + \beta_{Aj}X_{Aij} + r_{ij} \\
&\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
&\beta_{Tj} = \gamma_{T0} + u_{Tj} \\
&\beta_{Rj} = \gamma_{R0} + u_{Rj} \\
&\beta_{Mj} = \gamma_{M0} + u_{Mj} \\
&\beta_{Aj} = \gamma_{A0}
\end{aligned}
\qquad (19)
$$

The equation uses the same terms as the propensity score model but adds the regression coefficient for the treatment status, β_{Tj}, and the error term at the student-level, r_{ij}. The parameter values for the TE data generating model are as follows: $\gamma_{00} = 1.58$; $\gamma_{01} = -.025$; $\gamma_{T0} = -.40$; $\gamma_{R0} = .61$; $\gamma_{M0} = .24$; and $\gamma_{A0} = -.006$. The parameter value of $\gamma_{T0} = -.40$ is equivalent to a Cohen's *d* effect size for the treatment effect, of -0.6.

To reflect the empirical dataset, the data were generated to allow for correlations among the covariates and among the random slope variances of the student-level variables. At the school-level, the number of students retained in the prior school year (*W)* and the school means of kindergarten reading (X_R), kindergarten math (X_M), and age at kindergarten entry (X_A) and the school-level residuals for the PS model and the TE model were generated as random normal variables. The student-level covariates were also generated as random normal variables. The mean and covariance structure of all covariates and school-level residuals are shown in Table 1.

Table 1

Means, covariance and correlational structures of school- and student-level covariates and school-level residuals

School-level covariates and student-level covariates at the school-level (for one ICC condition)[a]				
	X_R school mean	X_M school mean	X_A school mean	W
Means[b]	0.46	0.45	66.12	1.71
Covariance				
X_R school mean	0.19			
X_M school mean	0.15	0.18		
X_A school mean	0.14	0.16	5.11	
W	-0.08	-0.14	0.03	5.15
Correlations				
X_R school mean	1.00			
X_M school mean	0.81	1.00		
X_A school mean	0.14	0.17	1.00	

W	-0.08	-0.15	0.01	1.00
School-level residuals in the PS model				
		u_{0j}	u_{Rj}	u_{Mj}
Means		0.00	0.00	0.00
Covariance				
u_{0j}		6.58		
u_{Rj}		0.32	4.15	
u_{Mj}		1.86	-2.13	2.93
Correlations				
u_{0j}		1.00		
u_{Rj}		0.06	1.00	
u_{Mj}		0.42	-0.61	1.00
School-level residuals in the TE model				
	u_{0j}	u_{Tj}	u_{Rj}	u_{Mj}
Means	0.00	0.00	0.00	0.00
Covariance				
u_{0j}	0.12			
u_{Tj}	-0.02	0.08		
u_{Rj}	0.01	0.02	0.02	
u_{Mj}	-0.01	0.01	-0.01	0.01
Correlations				
u_{0j}	1.00			
u_{Tj}	-0.21	1.00		
u_{Rj}	0.12	0.39	1.00	
u_{Mj}	-0.20	0.29	-0.71	1.00
Student-level covariates within schools[c]				
		X_R	X_M	X_A
Covariance				
X_R		0.47		
X_M		0.34	0.46	
X_A		0.49	0.69	16.55
Correlations				
X_R		1.00		

X_M	0.73	1.00	
X_A	0.18	0.25	1.00

Note. [a]To create four different conditions of the intracluster correlations, the covariance matrix was multiplied by .25, .5, 1, and 2.
[b]Means of the school-level means.
[c]The within-school means vary across schools as shown in the first covariance matrix of this table.

In both the PS and the TE models, data were generated so that the student-level covariates were centered at the school mean, and the school-level covariate was centered at the grand mean. This was consistent with how the empirical data were analyzed, based on the recommendations from methodological research on centering in multilevel models (Enders & Tofighi, 2007). The centering choice for the school-level covariate is straightforward since there is only the choice to center at the grand mean or to not center at all. Centering at the grand mean is more interpretable because the intercept represents the expected outcome when the covariate is equal to the mean across all of the clusters. For student-level covariates, the choice is more complex because one can center at the cluster mean or at the grand mean. Enders and Tofighi show that centering within clusters removes the between-cluster variation, which leads to more accurate estimates of slope variance. Because this simulation involved estimating slope variances of the student-level predictors, centering within clusters was more appropriate.

The simulation manipulated five factors. The number of individuals within a cluster and the intraclass correlation coefficients (ICCs) of the student-level covariates were between-cell factors, and the propensity score model, matching method, and balance measures were within-cell factors. The ICC for each student-level covariate was calculated as follows:

$$ICC = \frac{\sigma^2_{cluster}}{\sigma^2_{cluster} + \sigma^2_{student}} \quad (20)$$

Where $\sigma^2_{cluster}$ is variation explained by differences between clusters and $\sigma^2_{student}$ is variation explained by differences between students within clusters. Of the manipulated factors, the ICCs of the student-level covariates was the only factor that affected the data generation parameters. Specifically, the explained variance across clusters presented in the top matrix of Table 1 (the school-level covariates and student-level covariates at the school-level) varied across the four ICC conditions. Based on the covariance structure obtained from the ECLS-K dataset and presented in Table 1, the ICCs of kindergarten reading, math, and age at entry were calculated as .29, .28, and .24, respectively. To obtain the desired ICCs at three other levels, the full covariance matrix of the cluster-level covariates was multiplied by .25, .50, and 2. Table 2 illustrates the ICCs across the four conditions. The full set of variance/covariance parameters of these four conditions are reported in the appendix.

Table 2

Intracluster correlations of the unit-level covariates across four factor levels

Factor level	1	2	3	4
Ratio of the school-level covariance matrix in Table 1	0.25	0.50	1.00	2.00
Kindergarten reading achievement ICC	0.09	0.17	0.29	0.45
Kindergarten math achievement ICC	0.09	0.16	0.28	0.44
Age at kindergarten entry ICC	0.07	0.13	0.24	0.38
Average ICC across covariates	0.08	0.15	0.27	0.42

Note. ICC=intracluster correlation.

Data were generated with 500 replications within the cells of the study conditions described in the next section, which is a common number of replications in other simulations using multilevel PS methods (Arpino & Cannas, 2016; Li et al., 2013;

Rickles & Seltzer, 2014). To confirm that the number of replications was appropriate, the convergence of the simulation outcomes (selection of the correctly specified model and correlations between the balance measure and bias) was examined across the 500 replications. The results from this analysis are provided in Chapter 4.

3.2 Manipulated and Fixed Factors

As shown in Table 3, the simulation included both between-cell and within-cell factors. The between-cell factors included the ICCs of the student-level covariates and the cluster sizes, and the within-cell factors included the PS models, matching methods, balance measures, and the method of summarizing the balance measures across the covariates. The fixed factors included the coefficients in the PS model and the TE model (Table 1), and the number of clusters (50).

Table 3

Manipulated factors and levels

Manipulated factors	Levels
Between cell conditions	
Average ICCs of unit-level covariates	.08
	.15
	.27
	.42
Cluster sizes	10
	25
	100
	400
Within cell conditions	
Propensity score model	Correctly specified RIS model
	Over-parameterized, RIS model
	Under-parameterized, RI model
	Under-parameterized, SL model
	Under-parameterized, SL model without W_j
Matching method	Pooled
	Two-stage
	Within-cluster
Balance measures	Pooled ASB
	ASB
	Indicator of >.1
	Indicator of >.25
	Within-cluster ASB
	Mean across clusters
	Median across clusters
	Percentage of clusters >.1
	Percentage of clusters >.25
	Pooled variance ratio
	Variance ratio
	Indicator of <.5 or >2
	Within-cluster variance ratios
	Mean across clusters
	Median across clusters
	Percentage of clusters <.5 or >2
Summary of balance measures across covariates	Mean
	Weighted mean

Note. RIS=random intercepts and slopes; RI=random intercepts; SL=single-level; W_j=cluster-level covariate; ASB=absolute standardized bias.

3.2.1 Between cell conditions. Because the number of students within clusters and the ICCs of the student-level covariates have been shown to affect the modeling and matching steps, these were important factors to vary when investigating the diagnostics step. For example, Thoemmes and West (2011) found that when the ICCs of the unit-

level covariates were high (.5), the random intercepts and slopes PS model was preferred for reducing bias of the TE estimate, but when the ICCs were low, bias was low for all of the PS models tested. This means that it may be more difficult for balance measures to detect model misspecifications when the ICCs are low compared to when they are high, and for this reason, it was important to assess the ability of balance measures to detect model misspecifications with varying ICCs. Thoemmes and West (2011) considered ICCs of .05 and .5 for their unit-level covariates but warned that an ICC of .5 is higher than what one would expect in an applied study. For this reason, the simulation tested ICCs of .24, .28, and .29, which are the ICCs of the variables age at kindergarten entry, kindergarten math, and kindergarten reading, respectively, from the ECLS-K dataset, as shown in Table 2. The average ICCs in the other conditions are .08, .15, and .42, which are also more realistic than the ICCs tested by Thoemes and West.

Cluster sizes are another important consideration for selecting the modeling and matching approach in a multilevel study, and thus, were also important in selecting the diagnostic approach. For instance, Arpino and Cannas (2016) found that within-cluster matching resulted in greater bias of the TE estimate compared to pooled matching when clusters were smaller than 300, and the random intercepts and slopes model resulted in greater bias in the TE estimate than the random intercept only model when clusters were smaller than 20. Because the interaction between the cluster size and the matching and modeling method has an effect on the bias of the TE estimate, it was important to assess balance measures with different cluster sizes. It was also important to test a wide range of cluster sizes, because a wide range of cluster sizes are used in applied settings, depending on the context. One researcher may focus on the nested structure of students within

classrooms while another may focus on the nested structure of students within schools. To represent a range of contexts, this simulation used cluster sizes of 10, 25, 100, and 400.

3.2.2 Within cell conditions. Within cells, the simulation varied the PS models imposed on the data, matching methods used to equate the groups, and balance measures used to evaluate the equivalency across groups. The correctly specified RIS model used to generate propensity scores in Equation 18, was compared to four misspecifications that represent four common modeling errors that researchers could make. The first misspecified PS model is an over-parameterized (OP) model in which a student-level variable that has no relation to treatment selection, the outcome, or any predictors of treatment selection is included in the model. The modeled relation of the fourth student-level variable (X_4) to treatment status varied randomly across clusters and did not interact with other variables in the prediction of treatment status or the outcome. The remaining misspecifications are under-parameterized models that fail to include random components and/or variables. Each of these models is nested such that the random intercepts (RI) model is a reduced version of the RIS model, the single-level (SL) model is a reduced version of the RI model, and the model without cluster-level covariates (NoL2) is a reduced version of the SL model. Specifically, the RI excludes u_{Rj} and u_{Mj} from Equation 18, because the intercepts vary across clusters but the relations between the unit-level variables and treatment status does not. Researchers might select the RI model if they incorrectly assume that the multilevel setting is a broad rather than a narrow inference space. A further misspecification would be to use an SL model, treating the cluster-level variable, *W*, as if it were a unit-level variable. To demonstrate this error,

the SL model removes u_{Rj} and u_{Mj}, the random slopes, and u_{0j}, the random intercepts from Equation 18, making the intercepts fixed across clusters. Finally, the NoL2 model represents an SL model that fails to include any cluster-level predictors of treatment selection. This misspecification is achieved by removing u_{Rj}, u_{Mj}, u_{0j}, and $\gamma_{01}W_j$ from Equation 18.

The simulation also tested the three matching methods used for multilevel PS matching—pooled, two-stage, and within-cluster matching. Each method has been shown to be appropriate in narrow inference spaces when paired with an RIS, PS model; however, within-cluster matching can lead to biased TE estimates with small cluster sizes (Arpino & Cannas, 2016; Arpino & Maelli, 2011; Rickles & Seltzer, 2014). In the pooled matching condition, treated units were matched to control units with the closest PS estimate, regardless of cluster membership, while in the within-cluster matching condition, matches were restricted to control units in the same cluster. In the two-stage matching condition, a match was first attempted within the same cluster before moving to the pooled sample. The two-stage matching was not restricted to cluster groups as in Rickles and Seltzer but instead was open to the pooled sample as in Aprino and Cannas. All matching conditions were implemented with nearest neighbor matching with a caliper of .2 standard deviations of the PS with units matched in a random order. This matching method is common in applied studies and has been shown to be effective for removing selection bias (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1985; Thoemmes & Kim, 2011). In the within-cluster matching condition, if there were no individuals in the control group within the designated caliper width and the same cluster, then the treated unit was not matched. By contrast, in the two-stage matching condition, any treated

individuals that could not be matched within the designated caliper width and cluster were then matched to an individual within the caliper width but from another cluster.

The simulation tested both pooled and within-cluster forms of absolute standardized bias (ASB) and variance ratios for examining balance. As described previously, standardized bias is the difference in treatment and control means divided by the pooled treatment and control standard deviations (Equation 7), and is the most common metric for evaluating balance in PS-matched samples. In particular, ASB, the absolute form of standardized bias (Equation 13) is the most commonly reported balance measure in studies that have investigated multilevel PS methods (Arpino & Cannas, 2016; Arpino & Maelli, 2011; Rickles & Seltzer, 2014). Although it has not yet been adopted by many applied researchers, the variance ratio (VR, Equation 8) is recommended for assessing balance of the sample distribution (Austin, 2009). Understanding how to assess balance both in terms of the sample means and the distributions of the covariates is important for multilevel studies using PS methods.

The pooled balance measures were calculated for each covariate in the PS model of the full matched sample, ignoring cluster membership. To facilitate comparison of VRs across clusters and covariates, the VR was calculated so that the smaller variance was always the numerator and the larger variance was always the denominator. Additionally, binary (0/1) variables were created to indicate whether each pooled measure was above or below commonly accepted thresholds. For ASB, the thresholds were set to .1 and .25, which are commonly used to evaluate bias in a PS model (e.g., Harder et al. 2010; Normand et al., 2001). Based on the What Works Clearinghouse group design standards, an ASB of .1 on a pre-test measure would require the researcher

to make a covariate adjustment to the TE model in order to get a rating of "meets standards with reservations," and an ASB of .25 on a pre-test measure would automatically result in a rating of "does not meet standards" (U.S. Department of Education, 2017). For the pooled variance ratio, a binary variable indicated whether the variance ratio was below .5 or greater than 2, which suggest extreme differences in sample distributions (Rubin, 2001).

Each of the same statistics were calculated for the within-cluster balance measures as with the pooled balance measures (ASB, variance ratio, and threshold variables for all unit-level covariates), but were calculated separately for each cluster. To summarize the ASB and variance ratios across all clusters, the mean and median of the cluster-level balance measures were calculated. It was important to include the median as well as the mean as a measure of central tendency because it is less sensitive to outlying clusters. The binary threshold variables were reported as a percentage of clusters to show the extent of clusters with problematic levels of balance.

Once all of the balance measures were calculated for each covariate, they needed to be summarized into one metric that could be used for decision-making. Two approaches were tested: in the first approach, the researcher considers all covariates to be equally important and calculates the mean of the balance measure across covariates; in the second approach, the researcher applies weights to the covariates in terms of their influence on the outcome. Although the first approach is intuitive, Ho et al. (2007) recommended prioritizing the balance of covariates that more strongly influenced the outcome in the TE model. To do so, the weight of each covariate in the weighted mean was determined based on the influence of each covariate on the outcome in the real data.

Specifically, they were weighted according to the t-values for each covariate in the data generating TE model, where the weight of each covariate X_R, X_M, X_A, and W was equal to 63, 25, 6, and 5, respectively. The t-values show the strength of the covariate in predicting the outcome measure according to the t-distribution, and can be converted into different types of effect sizes (Durlak, 2009). It is worth noting that a researcher would not have these precise estimates, because the TE would not be estimated until after matching. Instead, the researcher would consult with prior research to determine the relative importance of each covariate in estimating the treatment effect. This weighted approach is similar to a balance measure proposed by Stuart, Lee, and Leacy (2013), the ASB of the prognostic score. A prognostic score is a single score that summarizes a person's likelihood of achieving a dichotomous outcome, such as passing a test or graduating from high school (Hansen, 2008). Although Hansen formally defined this term, Stuart et al. were the first to propose its use as a balance measure. In this context, Stuart et al. used both propensity scores and prognostic scores: first treatment units were matched to control units based on the propensity score; then, the ASB of the prognostic score was calculated to assess balance. Stuart et al. found that the ASB of the prognostic score had the highest correlation with bias in TE estimate compared to other balance measures and that it worked well in a variety of circumstances. This simulation could not use the prognostic score because of its use of a continuous outcome variable; however, the weighted mean as described above was a close proxy that could be tested with each of the balance measures, including the VR and indicators (ASB>.1, ASB>.25, and VR <.5 or >2).

3.2.3 Fixed factors. The parameters for both the PS and the TE data generating models remain constant across study conditions, except for the cluster-level covariate and the school means of the student-level covariates. These parameters vary in order to vary the ICCs of the unit-level covariates in the PS model. Because the focus of this simulation is on the identification of misspecifications of the PS model and not the TE model, the simulation does not include misspecifications of the TE model. Across all matched samples resulting from the study conditions, the TE estimate was calculated as a difference in treatment and control means. Furthermore, the simulation varied the number of units within each cluster, while holding the number of clusters constant at 50, the same number of clusters used in Rickles and Seltzer (2014). Fifty clusters may be a reasonable number faced by an applied researcher given that the empirical examples, undertaken with extant data included 859 and 29 clusters each.

3.3 Outcome Measures

The goal of the simulation is to assess the degree to which the tested balance summary measures can correctly select the correctly specified PS model and the degree to which they correlate with bias of the TE estimate. Before evaluating model misspecifications, it is important to determine the degree of misspecification of each alternative model. Several criteria can be used for model comparison from either a frequentist or a Bayesian perspective. From the frequentist perspective, a loglikelihood ratio test may be used to compare the goodness of fit between any two nested models, and one can calculate a p-value using the chi-squared distribution. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are two other common criteria for model comparison. Burnham and Anderson (2004) explain that the

choice about whether to use the AIC or BIC does not depend on the preference for using frequentist or Bayesian, since they can both be derived in either framework. Rather, the decision depends on one's philosophy about model selection. AIC assumes that there is no "true model" and that a different model may be selected with different sample sizes. By contrast, the BIC assumes that there is one "true model," a data generating model that is independent of the sample size. Given that the simulation is based on one data generating model, the philosophy of the BIC is better aligned with this simulation. However, Bayesian factors rely on the proper selection of priors. Alternative Bayesian criteria for model comparisons including fractional Bayes factors (O'Hagan, 1995) and intrinsic Bayes factors (Berger & Pericchi, 1996) have proposed ways to address this problem. But a remaining concern of the BIC is that it tends to produce biased results from the selected model when the sample size is small (Burnham & Anderson). Because the small cluster sizes tested in the simulation, AIC is more appropriate for assessing model fit. Therefore, the fit of the models were compared using likelihood ratio tests and AIC to determine the degree of misspecification for each alternative model.

All balance measures were compared in terms of their ability to select the correctly specified PS model (Equation 18). If a balance measure is effective at detecting bias, then it will indicate poorer levels of balance (higher ASBs and variance ratios further from 1) for the matched samples resulting from the misspecified PS models than for the matched samples resulting from the correctly specified PS model. By contrast, if a balance measure is ineffective at detecting bias, then it might incorrectly indicate that the samples resulting from the misspecified PS models are well balanced. In a set of Monte Carlo simulations, Austin (2009) compared one condition in which the PS model was

correctly specified to another condition in which a covariate was missing from the PS model. For each covariate, they calculated both the ASB and the variance ratio. They concluded that the variance ratio was better at detecting model misspecifications because the ASBs were nearly the same for the correctly specified and misspecified model but the variance ratios showed poorer balance for the misspecified model.

The current study expanded the approach of Austin (2009) by simulating the process that applied researchers would undertake when selecting a PS model. Applied researchers implementing the diagnostic step in a multilevel PS analysis would consider a few different specifications of the PS model and would then select the one resulting in the most balanced samples. To quantify this diagnostic approach, the selected "best" model was recorded for each balance measure, assuming that a researcher would select the model resulting in the greatest balance, as measured as the average balance across the covariates. Whenever there was an exact tie in balance between models, none was selected. Then, across the 500 replications, the selection of the best model for each balance measure was tallied and the percentage in which each model was selected was calculated. If the model was selected for more than 50 percent of the replications, then it was selected as the "winner," and if the model was selected for more than 75 percent of replications, it was selected as the "clear winner."

Effective balance measures should not only be capable of selecting the correctly specified PS model but should also predict bias in TE estimates. As such, a correlation between each pooled and within-cluster balance measure and bias of the TE estimate was calculated. With each replication, the simulation calculated balance statistics for the matched sample and an estimate of the TE. Bias of the TE was calculated for each

replication by subtracting the TE estimate from the true TE (Equation 19). Once all replications were completed, a correlation coefficient between the balance measure and bias across replications was calculated. Balance measures that are most effective should have higher correlations with absolute bias of the TE estimate. This approach was used to evaluate balance measures in other studies of PS methods (Belitser et al., 2011; Stuart et al., 2013).

Absolute bias was calculated rather than relative bias. In the case of measuring the correlation between balance measures and bias, absolute bias is more appropriate because greater levels of imbalance should lead to greater levels of bias. This would not be the case if the direction of bias was recorded.

3.4 Software

Data generation, matching, and analysis were carried out in SAS software, Version 9.4 for Windows[2]. The GMatch SAS macro developed by Brad Hammill (2015) was adapted for the multilevel matching procedures used in this study.

3.5 Summary of Simulation Procedures

This chapter described the simulation methods used to address the study's research questions: 1) "which balance measures are best at identifying the correctly specified PS model?", and 2) "which balance measures are most strongly correlated with bias in the TE estimates?" Figure 4 summarizes the steps of the simulation in the form of a flowchart. First, data were generated according to the specified PS and TE models (Equations 18 and 19) and the covariance structures of the variables and residuals (Table

[2]Copyright © 2013 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

1) for each of the 16 combinations of ICC and cluster size. Next, the five PS models were run on the 16 different datasets, and separate propensity scores were saved from each model. Then for each dataset and propensity score model, the treatment and control groups were matched in three different ways: pooled matching, two-stage matching, and within-cluster matching. Pooled and within-cluster balance measures and TE estimate bias were then calculated for each of the 240 matched datasets. As a reminder, TE estimates were calculated as the difference in treatment and control means in the matched sample. The within-cluster balance measures were summarized both as a mean and as a median across all of the clusters, and all balance measures were summarized across the covariates as a simple mean and as a weighted mean. These procedures were completed 500 times. Finally, the percentage of replications in which the RIS model was selected and the correlation between TE estimate bias and balance was calculated for each condition and balance measure. The results are reported in the next chapter.

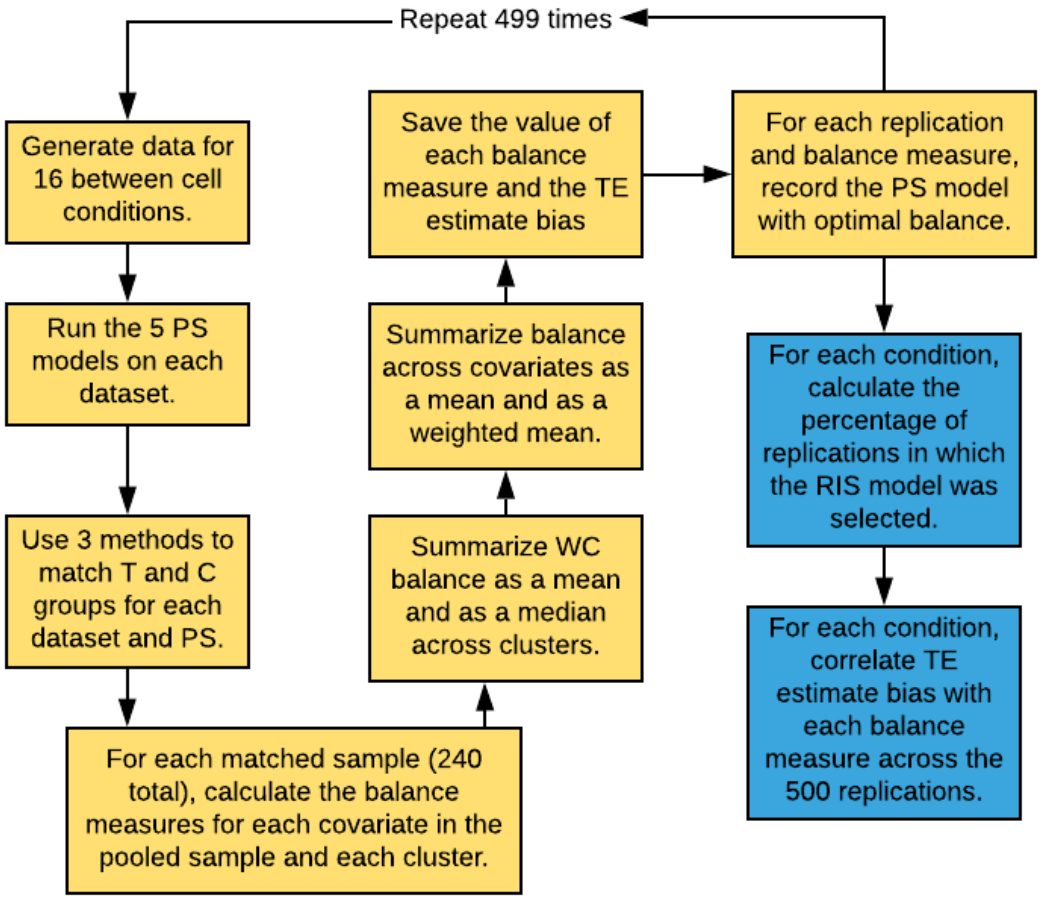

Figure 4. Flow of simulation procedures from data generation to outcome estimation. PS=propensity score; T=treatment; C=control; WC=within-cluster; TE=treatment effect.

Chapter 4. Simulation Results

The previous chapter described the methods used to answer two main types of research questions: 1) the balance measures that can properly identify the correctly specified PS model and 2) the balance measures that are most correlated with bias in the TE estimate. This chapter provides the results from the Monte Carlo simulation that answer these questions. Before discussing the results on the outcome measures, the chapter describes the convergence of the simulation outcomes across the 500 replications as evidence of the reliability of the results presented throughout the chapter. For context,

it also provides descriptive information about the TE estimate bias across the study conditions (ICCs of the student-level covariates, cluster sizes, PS models, and matching methods). It then describes the results pertaining to each research question by summarizing the results across and between the study conditions. The chapter provides tables and figures to illustrate the major findings; the comprehensive results for each condition of the simulation are presented in tables and figures in the appendix.

4.1 Convergence

The outcome measures successfully converged prior to the 500th replication. For the outcome of the percentage selection of the correctly specified model, convergence was defined as a change of no more than 1 percentile point from one replication to the next. On average across conditions, the percentage of replications in which the mean pooled ASB selected the correctly specified model converged within 52 replications, and the most replications required for any condition to converge was 94. For the correlation outcome, convergence was defined as a change of no more than .01 from one replication to the next. On average across conditions, the correlation between the mean pooled ASB and TE estimate bias converged within 313 replications. The number of replications required for convergence of the correlation outcome varied according to the cluster size, model, and matching method. In general, the within-cluster matching approach required more replications for convergence compared to the pooled or two-stage matching; the single-level models required more replications than the RIS models; and the largest cluster size (400) required more replications than the smaller cluster sizes. Just 2 of the 240 conditions did not meet the convergence criteria within 500 replications, but graphical inspection of these conditions revealed there was no need for concern for

interpreting the results. Figure 5 demonstrates the convergence of the correlation between TE estimate bias and each of the balance measures for one of these two conditions, in which the ICC was high, cluster size was 10, PS model was RIS, and matching method was WC. Each color represents a different balance measure. The horizontal lines represent the estimate of the correlation parameter across the 500 replications. The jagged lines close to the Y-axis represent the large changes in estimates from one replication to the next during the first few replications but gradually become smoother as the simulation progresses. They are relatively smooth by 150 replications, but there are some jags in the lines until approximately the 450th replication. However, all correlations appear stable by the 500th replication.

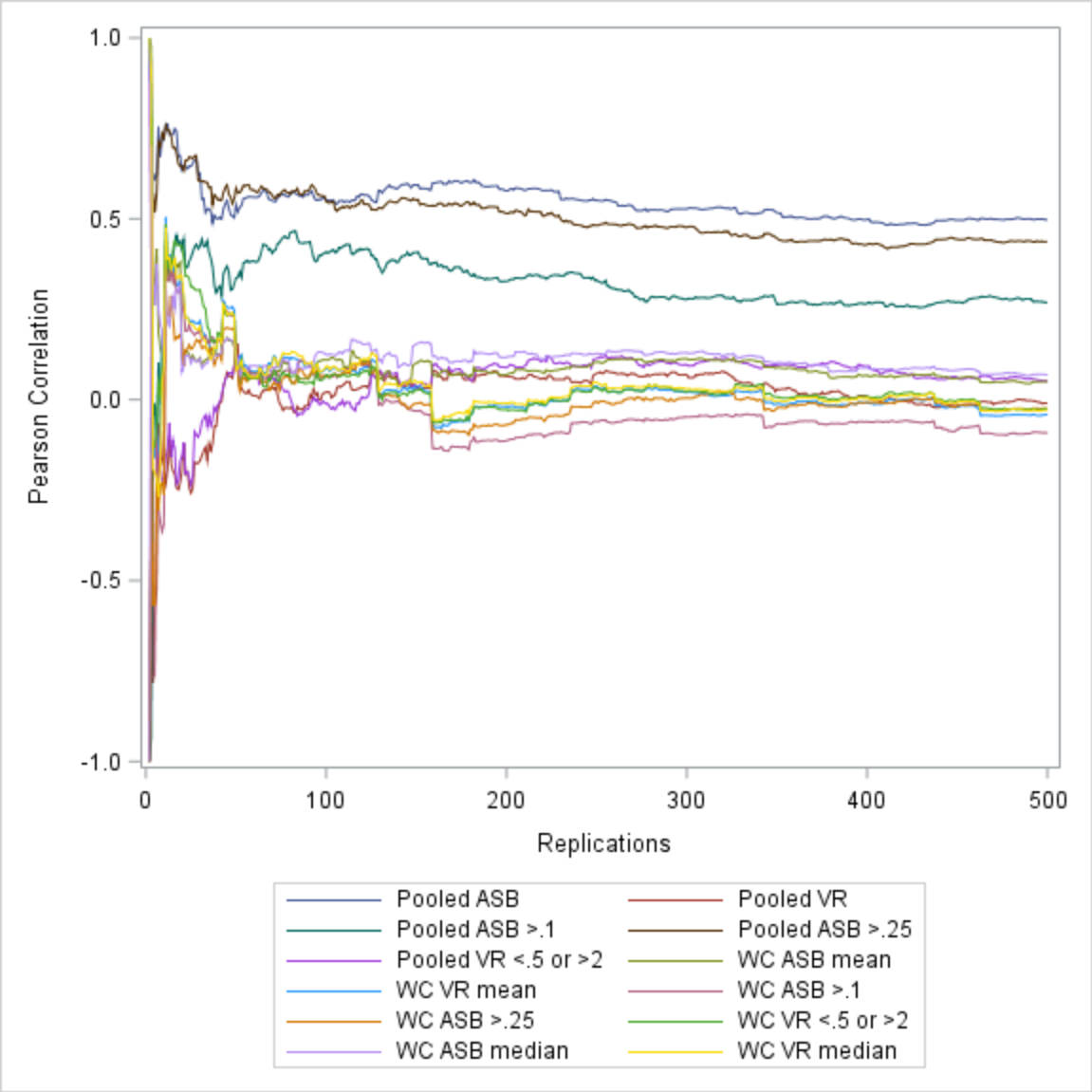

Figure 5. Convergence of the correlation between absolute treatment effect estimate bias and each of the balance measures in the condition where the average intracluster correlation is high (.42), there are 10 units per cluster, the propensity score model has random intercepts and slopes, and matching is conducted within clusters. The correlation between treatment effect estimate and each balance measure is represented by a different color, where ASB=absolute standardized bias, VR=variance ratio, Pooled=pooled balance measure, and WC=within-cluster balance measure. All balance measures are summarized across covariates as an equally-weighted mean; the convergence pattern looks the same for the balance measures summarized as unequally-weighted means.

4.2 Bias of the Treatment Effect Estimates

Before interpreting the results from the two research questions, it is important to first understand the values of the TE estimate bias across the simulation conditions. In

each condition, TE estimate bias was calculated as the absolute difference in the TE estimate and the TE parameter specified in the data generation. Without any type of matching, TE estimate bias was .54 on average across conditions and replications, but was reduced to .07 after matching, on average across ICCs, cluster sizes, PS models, and matching conditions. As shown in Table 4, bias was lowest for within-cluster matching (.04) compared to two-stage (.08) and pooled matching (.09). As expected, bias was reduced as the cluster size increased, from .11 for cluster sizes of 10, to .04 for cluster sizes of 400. However, the average bias for cluster sizes of 100 (.05) was just slightly greater than the average bias for cluster sizes of 400. Reviewing the results according to PS model, the RI model had the lowest bias on average across conditions (.04), followed by the RIS and OP models, which had almost the same level of bias (.07), followed by the SL model (.08), and finally, the model with no level-two covariates (.09). Varying the ICCs of the student-level covariates had very little effect on TE estimate bias but in general, across conditions, larger ICCs resulted in greater levels of bias. The study may not have seen as much of a difference in results according to ICC because the levels were based on real data and were more similar to one another than in Thoemmes and West (2011).

Table 4

Mean treatment effect estimate bias by ICC, cluster size, matching method, and model

Factor	Level	Mean and standard deviation of treatment effect estimate bias
ICC (mean across student-level covariates)	.08	.067 (.057)
	.15	.069 (.057)
	.27	.070 (.056)
	.42	.074 (.058)
Cluster size	10	.114 (.068)
	25	.075 (.044)
	100	.049 (.040)
	400	.042 (.042)
Model	RIS	.070 (.067)
	OP	.070 (.067)
	RI	.044 (.042)
	SL	.077 (.044)
	NoL2	.089 (.051)
Matching method	Pooled	.087 (.058)
	Two-stage	.082 (.052)
	Within cluster	.040 (.049)
All		.070 (.057)

Note. RIS=random intercept and slopes; OP=over-parameterized RIS model; RI=random intercept; SL=single level; NoL2=single level with no school-level covariates.

TE bias means for all conditions across the 500 replications are reported in the appendix.

It is also worth noting the interactions between PS models, matching methods, and cluster sizes on the TE estimate bias. As shown in Figure 6, bias was lowest for the RI model when pooled or two-stage matching was used, but bias was lowest for the SL model when within-cluster matching was used. Overall, the lowest level of bias was for the SL model paired with within-cluster matching (.026), and the highest level of bias was for the PS model without cluster-level covariates paired with two-stage matching. These results suggest that within-cluster matching can better control for variation between clusters than multilevel modeling. However, in the absence of within-cluster matching, the multilevel PS models can also reduce bias.

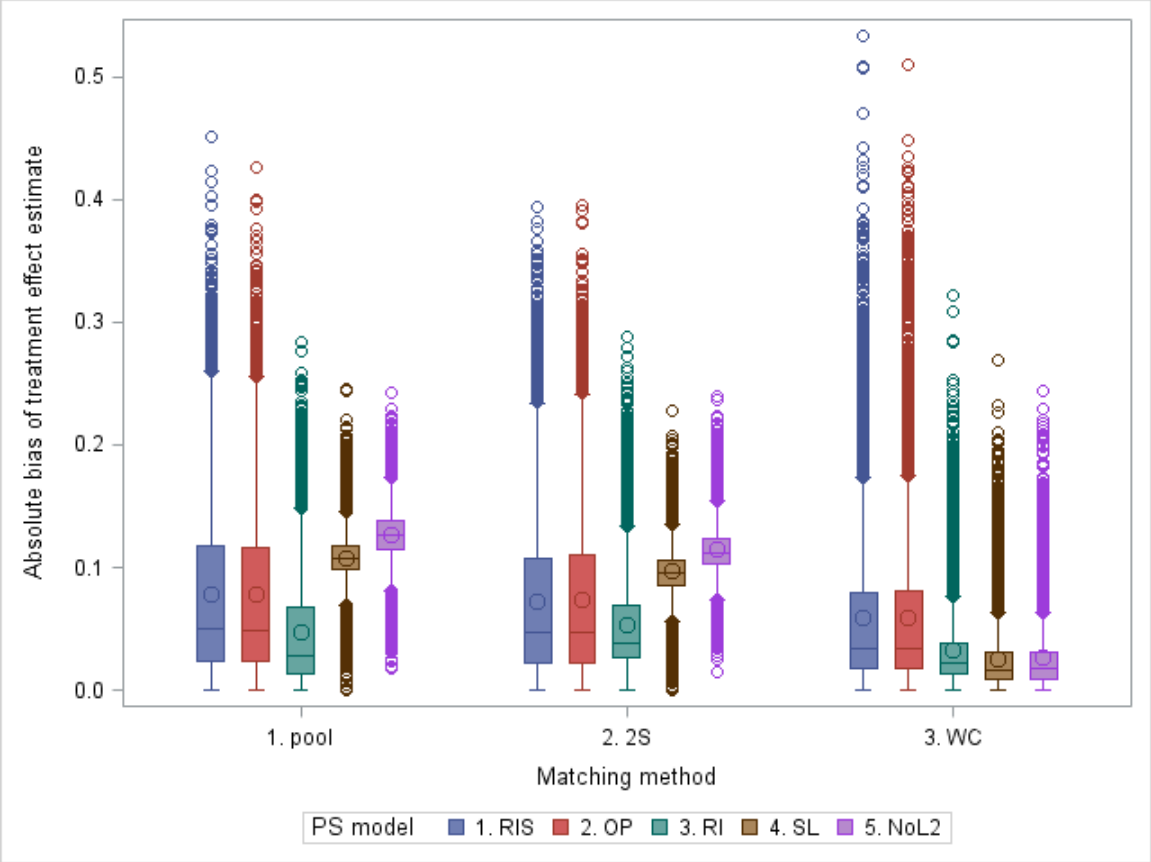

Figure 6. Absolute bias of the treatment effect estimate by matching method and propensity score model. Pool=pooled matching; 2S= two-stage matching; WC=within-cluster matching; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=single-level model without cluster-level covariates.

As shown in Figure 7, bias was greatest when clusters were small, but the difference was more pronounced for the more complex PS models (RIS and OP). For example, for the SL model, the average bias was between .088 and .071 across the tested cluster sizes. By contrast, the average bias for the RIS model was .018 with the cluster size of 400 but was .147 with the cluster size of 10. For the smallest cluster size (10), bias was lowest for SL model, but for cluster sizes of 25 and 100, bias was lowest for the RI model. For the largest cluster size (400) bias was lowest for RIS model.

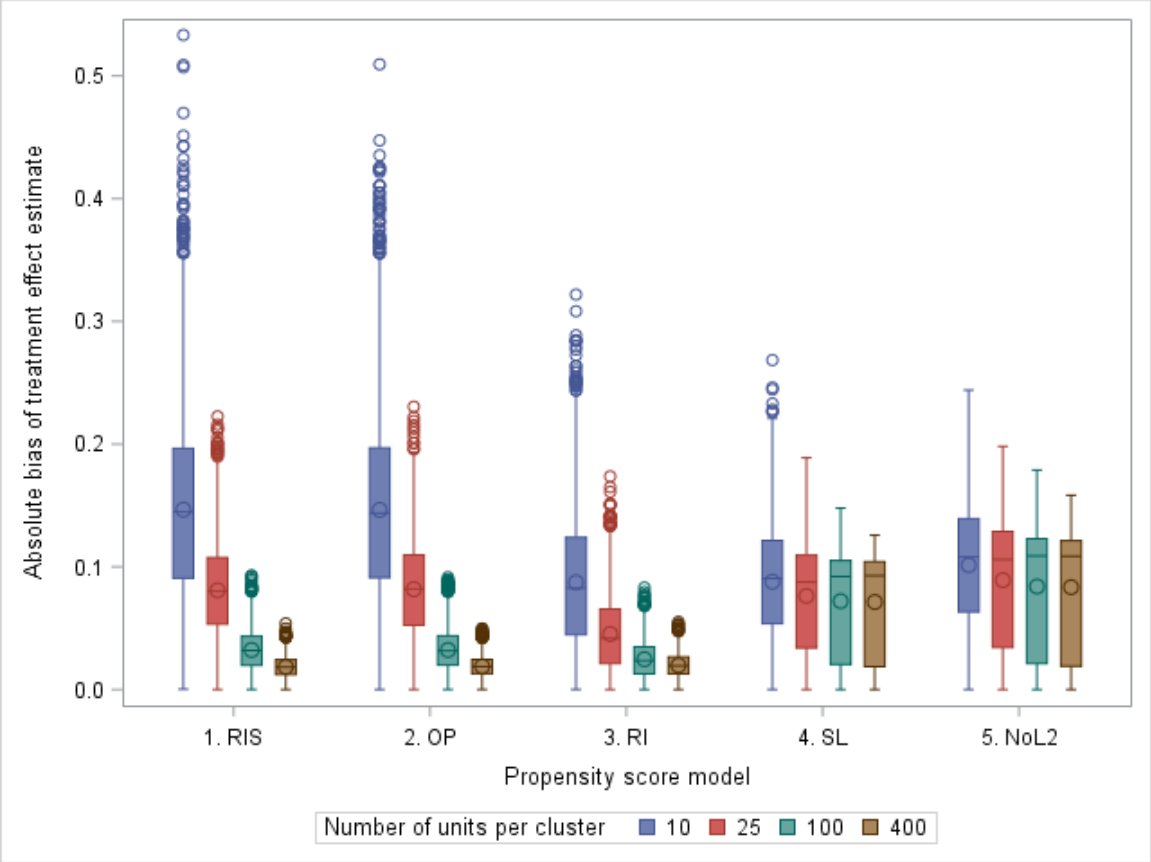

Figure 7. Absolute bias of the treatment effect estimate by propensity score model and cluster size. RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=single-level model without cluster-level covariates.

All together, these results indicate that with certain cluster sizes and matching methods, selecting the correctly specified PS model (the RIS model) can result in greater levels of TE estimate bias than a misspecified PS model. This finding has important implications for interpreting differences between the two simulation outcomes presented later in the chapter—the ability of each balance measure to select the correctly specified PS model, and correlation between the balance measure and TE estimate bias. Because the correctly specified PS model and the model that leads to the greatest reduction in TE estimate bias are not necessarily the same, there may be different balance measures that are optimal for each outcome.

4.3 Research Question 1: Which Balance Measures Identified the Correctly Specified Model?

Before interpreting the ability of the balance measures to identify the correctly specified model, it is important to put the differences between the model specifications into context. To do so, the -2loglikelihood and AIC model fit indices were recorded for each matched dataset. The -2loglikelihood and AIC were then averaged across replications and simulation conditions. Across cluster sizes and ICCs, the model fit indices followed the expected pattern based on the differences between each model and the data-generating model. Table 5 shows the AIC and -2loglikelihood values and the results from the likelihood ratio chi-square test of model differences for one replication; however, the pattern was similar across all of the replications. As shown in Table 5, the best fitting PS model was the RIS model, the model used to generate the data. The OP model had a slightly worse model fit according to the AIC, and although the -2loglikelihood value was lower, the likelihood ratio chi-square test revealed that it was not significantly different from the RIS model. These models were expected to have similar fit given that the OP model was the same as the RIS model but included one extraneous covariate.

The remaining models had progressively worse fit in the expected order: the RI model had poorer fit than the RIS model, the SL model had poorer fit than the RI model, and the SL model with no level-two covariates had poorer fit than the SL model. Each of these steps in reducing the RIS model to the model with no level two covariates were statistically significant at the $p < .0001$ level. However, the greatest change in model fit was between the RI model and the SL model.

Table 5

Model fit for one replication with the average ICC among the individual-level covariates=.42 and 10 units per cluster

Model	Number of parameters	AIC	-2loglikelihood	χ^2 difference from next reduced model
OP	9	423.60	407.60	0.02
RIS	8	421.62	407.62	19.11**
RI	6	438.73	426.73	131.79**
SL	5	568.52	558.52	25.75**
No L2	4	592.27	584.27	NA

**p<.0001

Note. OP=over-parameterized model; RIS=random intercept and slopes; RI=random intercept; SL=single level; No L2=single level with no school-level covariates. NA= not applicable because it is the most reduced model.

This pattern was consistent across the replications, but as shown in Table 6, there were differences in the percentage of significant likelihood ratio chi-square tests according to cluster size. Across all cluster sizes and replications, the difference in fit between the RIS model and the OP model was only significant for approximately 5% of the replications, and the differences between the RI and RIS model, the SL and RI model, and the NoL2 and the SL model were almost always significant (for 93%, 100%, and 94% of replications, respectively). However, the rates for significant differences between models were lower when there were 10 units per cluster; in this case, the differences between the RI and RIS model was significant for 72% of replications and the difference between the NoL2 and the SL model was significant for 80% of replications. This suggests that there are likely to be differences in the probability of selecting the correct model according to cluster size.

Table 6

Percentage of replications for which the likelihood ratio chi-square test was significant, by cluster size

Model comparison	Number of units per cluster 10	25	100	400	All
RIS vs. OP	5.0%	5.0%	5.3%	5.1%	5.1%
RI vs. RIS	72.2%	100.0%	100.0%	100.0%	93.1%
SL vs. RI	100.0%	100.0%	100.0%	100.0%	100.0%
NoL2 vs. SL	79.8%	97.3%	100.0%	100.0%	94.3%

Note. OP=over-parameterized model; RIS=random intercept and slopes; RI=random intercept; SL=single level; No L2=single level with no school-level covariates. NA= not applicable because it is the most reduced model.

With an understanding of the differences in fit between the models, we can now turn to interpreting the ability of each balance measure to select the RIS model, which was used to generate the PS data. As a reminder, within each condition and replication, balance of the matched sample was assessed using a variety of measures. Then, the PS model resulting in the greatest covariate balance between treatment and control groups was selected. If there was not a single best PS model for achieving balance, then none were selected. After completing the 500 replications, the percentage of replications in which the RIS model was selected was calculated. If a model was selected for at least 50% of the replications, it was considered the winning model.

Overall, across all conditions, the within-cluster ASB balance measures were most effective for identifying the correctly specified model, as shown in Table 7. This makes sense given that the correctly specified model included random intercepts and slopes, which meant that balance should differ across clusters. Therefore, the within-cluster balance measures would be better able to capture these within-cluster imbalances. Of the within-cluster ASB measures, the median of the cluster-level ASBs had slightly

higher rates of selecting the RIS model than the mean of the cluster-level ASBs, and both of these measures had higher rates of selecting the RIS model than the ASB indicators with thresholds of .1 and .25 (the percentage of clusters with an ASB >.1, and the percentage of clusters with an ASB > .25, respectively). The within-cluster ASB measures had higher rates of identifying the correctly specified PS model than the within-cluster VR measures. The rates for selecting the correctly specified model were slightly higher for the balance measures that summarized the covariates according to a weighted mean (based on the strength of the covariate's relation to the outcome variable) rather than an equally weighted mean. This was true for both within-cluster and pooled balance measures.

Table 7

Percentage of replications in which the RIS model was selected on average across ICCs, cluster sizes, and matching methods

Type of balance measure		Summarization of covariates	Percentage
Pooled	ASB	Mean	9.1%
		Weighted mean	10.7%
	ASB>.1	Mean	0.4%
		Weighted mean	0.7%
	ASB>.25	Mean	0.1%
		Weighted mean	0.1%
	VR	Mean	25.2%
		Weighted mean	23.6%
	VR<.5 or VR>2	Mean	0.1%
		Weighted mean	0.1%
Within-cluster	ASB mean	Mean	43.8%
		Weighted mean	46.1%
	ASB median	Mean	45.0%
		Weighted mean	46.5%
	ASB >.1 percentage	Mean	39.5%
		Weighted mean	42.3%
	ASB >.25 percentage	Mean	40.9%
		Weighted mean	44.3%
	VR mean	Mean	6.5%
		Weighted mean	9.0%
	VR median	Mean	7.8%
		Weighted mean	10.5%
	VR<.5 or VR>2 percentage	Mean	7.7%
		Weighted mean	10.5%

Note. ASB=absolute standardized bias. VR=variance ratio, The percentages for each condition are provided in the appendix.

The pattern of the types of balance measures that were best able to select the correctly specified model were not the same for the within-cluster and pooled balance measures. In the case of pooled balance measures, the VR was better able to select the RIS model than the ASB. Another key difference was that the indicators of ASB>.1, ASB>.25, and VR<.5 or >2 were not capable of selecting the RIS model in the pooled sample. This was because in nearly all of the matched samples the covariates rarely had ASBs above .1 or VRs below .5 or above 2. This resulted in values of 0 for the indicators ASB>.1, ASB>.25, and VR<.5 or >2 in the pooled sample for more than one model,

which meant that there was an exact tie between multiple PS models in a given replication. According to the study design, whenever there was an exact tie between multiple PS models, no model was selected.

Some balance measures differed in their performance in selecting the RIS model according to cluster size and matching method. Figure 8 shows the average rates of selecting the correctly specified model for the balance measures by each combination of cluster size and matching method. The figure shows the rates for the balance measures summarized across covariates as a simple mean, but the pattern is the same for the balance measures summarized according to the unequally weighted mean. As shown in Figure 8, across all matching methods and cluster sizes, the within-cluster ASB mean and median were most likely to select the RIS model. The rates for the pooled VR and ASB indicators (ASB >.1, ASB>.25, and VR <.5 or >2) were also consistent across all matching methods and cluster sizes, with rates near 0%. By contrast, the performance of the pooled ASB and VR varied according to matching method and cluster size. The pooled ASB had rates of selecting the RIS model below 10% for cluster sizes of 10, 25, and 100 units each, but for clusters of 400 units each, the rates of selecting the RIS model were much higher. Specifically, with a cluster size of 400, the rate of selecting the RIS model was 21% with pooled matching, 28% with two-stage matching, and 40% with within-cluster matching. The ability of the pooled VR also increased with cluster size, ranging from an average success rate of 10% for cluster sizes of 10 to an average success rate of 47% for cluster sizes of 400. For cluster sizes of 100 and 400, the rates of selecting the correct model also varied according to matching method with the highest rates of success with pooled matching. For example, with a cluster size of 100, the

percent that the pooled VR selected the correct model was 34% for pooled matching, 29% for two-stage matching, and 24% for within-cluster matching.

The different success rates of the pooled balance measures according to cluster size and matching method make sense in the context of the TE estimate bias described earlier in the chapter. As a reminder, the TE estimate bias was calculated as a pooled measure across all clusters. With small cluster sizes of 10 and 25, the TE estimate bias is lowest for the single-level model; therefore, it is more likely that a pooled balance measure will select the SL model than the RIS model. As the cluster size increases, it is more likely that the model that will reduce pooled TE estimate bias is the RIS model, so the selection of the RIS model based on a pooled balance measure is more likely. Similarly, TE estimate bias tended to be lower for the RIS model when within-cluster matching was used instead of another matching method, which may have also contributed to the differences in the selection rates for the pooled balance measures across the types of matching methods.

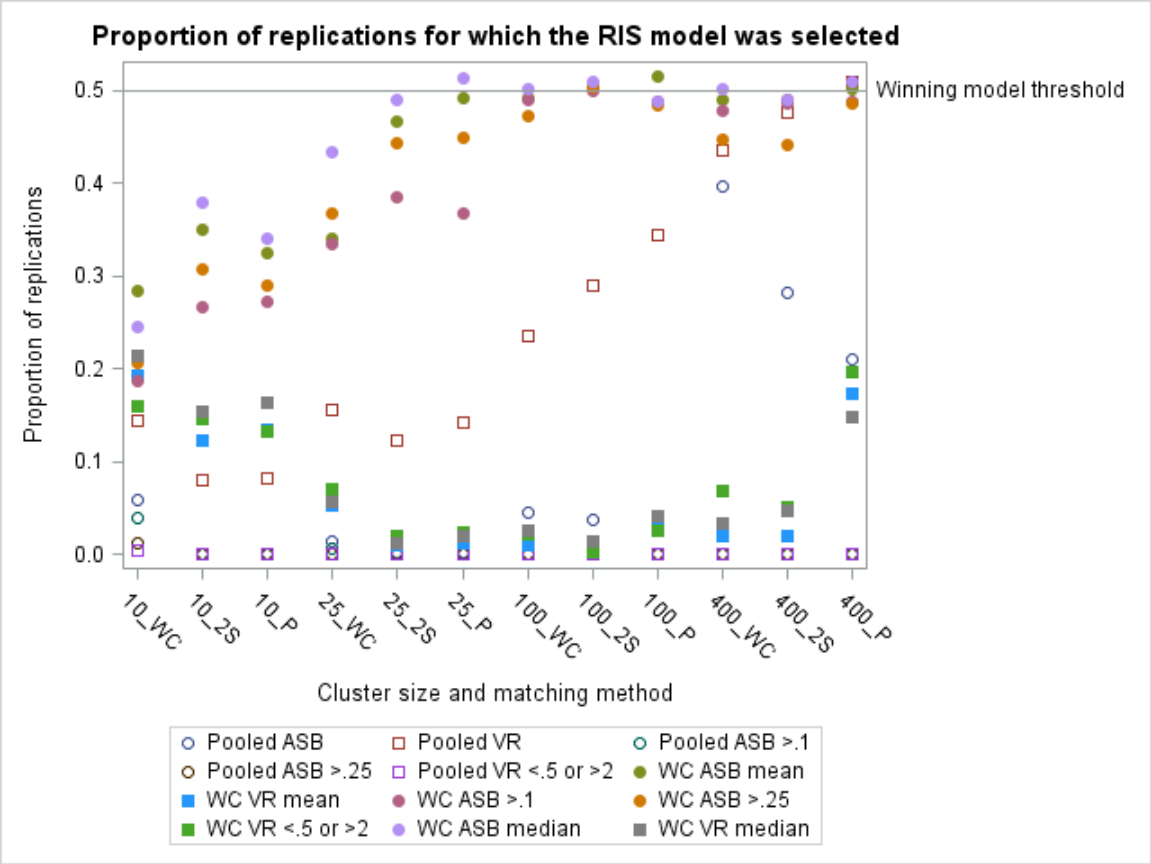

Figure 8. Proportion of replications for which the random intercepts and slopes (RIS) model was selected, by cluster size and matching method (WC=within-cluster matching, 2S=two-stage matching, P=pooled matching). Each circle or square represents a different balance measure (ASB=absolute standardized bias, VR=variance ratio, WC=within-cluster balance measure). This figure presents the mean of each balance measure across all covariates.

These results may seem discouraging; only a few of the best performing balance measures managed to select the correct model for about 50% of the replications. However, the descriptive and model fit information earlier in the chapter can help to explain this. An examination of model fit indicated that the over-parameterized model had nearly the same model fit as the RIS model. After all, the over-parameterized model was also an RIS model but happened to include an unnecessary student-level covariate. Furthermore, across conditions, the treatment effect bias for the over-parameterized model was approximately the same as the treatment effect bias for the correctly specified

RIS model. This indicates that there would be no problematic consequences for including an additional covariate in the PS model that is unrelated to treatment status or the outcome. Given that the model fit and resulting TE estimate bias from the two models were approximately the same, one could then define selecting the correct model as selecting either of the two RIS models.

By redefining the outcome in terms of selecting either of the two RIS models, one can see the same pattern among the balance measures but with much more promising results. As shown in Figure 9, many of the within-cluster ASB measures for cluster sizes of at least 25 have a perfect ability to select an RIS model across 500 replications. These measures were previously hovering close to 50% because they had an equal probability of selecting the RIS model or the OP model, since they were nearly equivalent. With the redefined outcome, even the condition with a cluster size of 10 and within-cluster matching has greater than a 50% chance of selecting one of the two correct models with a within-cluster ASB balance measure.

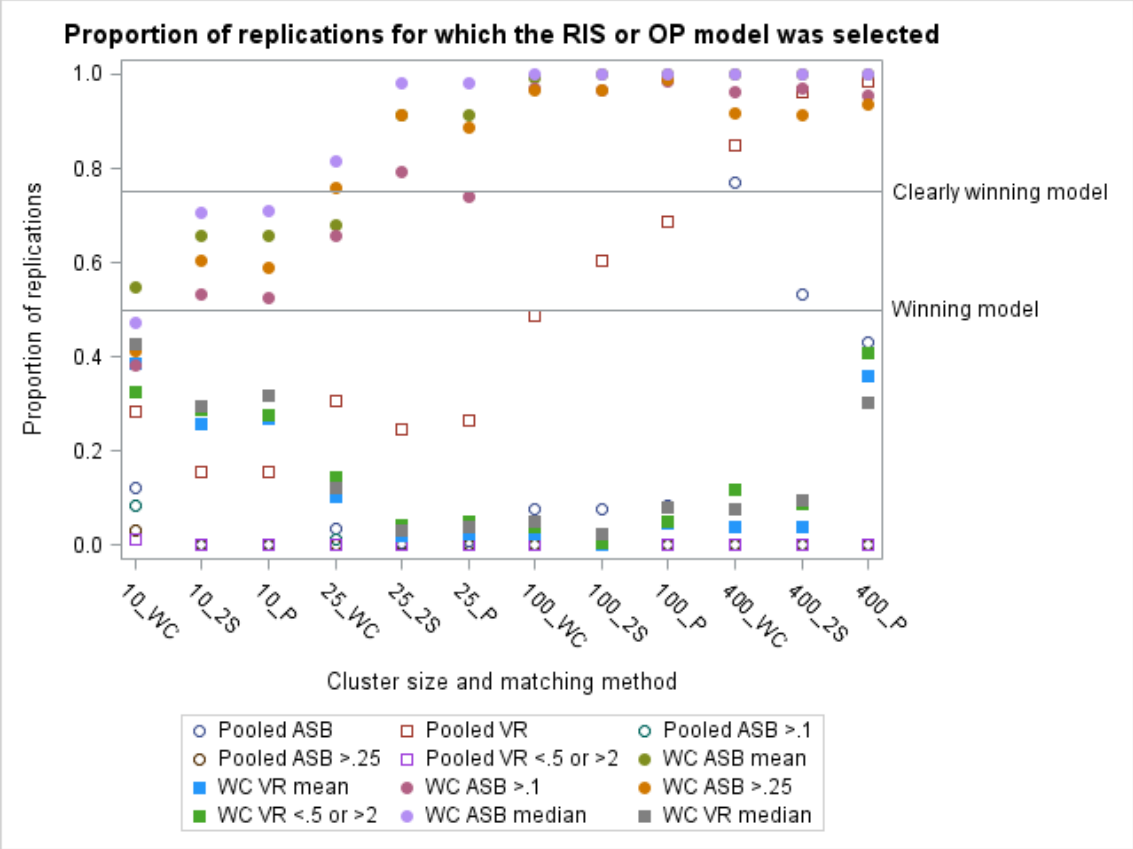

Figure 9. Proportion of replications for which the random intercepts and slopes (RIS) model or the over-parameterized (OP) model was selected, by cluster size and matching method (wc=within-cluster matching, 2s=two-stage matching, p=pooled matching). Each circle or square represents a different balance measure (ASB=absolute standardized bias, VR=variance ratio, WC=within-cluster balance measure). This figure presents the mean of each balance measure across all covariates.

4.4 Research Question 2: Which Balance Measures Are Most Strongly Correlated with Bias in the TE Estimate?

The balance measures were also evaluated based on the strength of correlation with TE estimate bias. Balance measures that are effective should be strongly, and positively, correlated with TE estimate bias so that high levels of imbalance should indicate that corrections should be made to the PS model. Likewise, low levels of imbalance should indicate that the researcher should proceed with the PS model and expect low levels of TE estimate bias. Before reviewing these results, it is important to recall from earlier in the chapter that the PS models that resulted in the lowest levels of

bias were not necessarily those with the best model fit. As shown in Tables 4 and 5, the RI model tended to reduce the TE estimate bias the most (Table 4), even though the RIS model had the best fit (Table 5). This suggests that the balance measures that are best at selecting the RIS model may be different from the balance measures that are best at predicting TE estimate bias. Given that the goal of PS matching is to achieve balanced samples that will reduce TE estimate bias, researchers should prioritize selecting the PS model that will reduce TE estimate bias over selecting the best fitting PS model, all else being equal. This section describes the results in terms of the correlations between the balance measures and TE estimate bias, which are the basis of recommendations for researchers presented in the empirical illustrations in Chapter 5.

Aggregated across all conditions, the pooled ASB measures had the strongest correlations, as shown in Table 8. On average across conditions, the correlation between TE estimate bias and the pooled ASB, summarized as an equally weighted mean across covariates, was .27; the correlation between the TE estimate bias and the pooled ASB, summarized as an unequally weighted mean, was .35. Based on the strength of the correlation, these two measures far outperformed all of the other measures, including the pooled VR measures and all of the within-cluster measures. On average across conditions, the pooled VR balance measures and the within-cluster measures each had correlations with TE estimate bias close to 0. As described previously, the pooled balance indicators (ASB >.1, ASB >.25, and VR<.5 or >2) yielded little variation, because across the pooled samples, there were rarely any covariates with values meeting those thresholds. For example, in many conditions the mean and standard deviation for ASB >.25 was 0 across the 500 replications, which meant that a correlation coefficient could

not be calculated. For this reason, the results for the pooled indicator balance measures are not included in the results for the remainder of the chapter. However, they are provided in the appendix.

Table 8

Pearson correlations between TE estimate bias and the balance measures

Type of balance measure		Summarization of covariates	Pearson correlation
Pooled	ASB	Mean	.27
		Weighted mean	.35
	VR	Mean	-.01
		Weighted mean	.00
Within-cluster	ASB mean	Mean	.03
		Weighted mean	.05
	ASB median	Mean	.05
		Weighted mean	.07
	ASB >.1 percentage	Mean	.02
		Weighted mean	.04
	ASB >.25 percentage	Mean	.04
		Weighted mean	.06
	VR mean	Mean	.01
		Weighted mean	.02
	VR median	Mean	.01
		Weighted mean	.01
	VR<.5 or VR>2 percentage	Mean	.01
		Weighted mean	.01

Note. ASB=absolute standardized bias. VR=variance ratio. The correlations for each condition are provided in the appendix.

The pooled ASB had the strongest correlation with TE estimate bias, on average, across all ICCs, PS models, and matching methods tested. As shown in Figure 10, the pooled ASB, weighted according to the strength of its relation to the outcome, typically performed better than the pooled ASB, summarized as an equally weighted mean across covariates. The exception to this pattern was for within-cluster matching; when within-cluster matching was used, the pooled ASB, summarized as a mean across covariates (rather than the unequally weighted mean), was more strongly correlated with TE

estimate bias; however, the correlations were similar (r= .23 for the simple mean and r=.21 for the unequally weighted mean).

Figure 10 also shows that there were differences in the preferred balance measure according to cluster size. For cluster sizes of 10, 25, and 100, the pooled ASB (with unequal weights across covariates) was most highly correlated with TE estimate bias, on average across ICCs, PS models, and matching methods (r=.51, .51, and .37, respectively). However, for the cluster size of 400, nearly all balance measures had correlation coefficients near 0 because the TE estimates were also near 0. On average across ICCs, PS models, and matching methods, when the clusters had 400 units each, the highest correlation was with the within-cluster ASB mean, but this correlation was only .04.

The figure also shows interactions between the cluster size and matching method on the strength of the correlations between the balance measure and TE estimate bias. When within-cluster matching was used, the pooled ASB as an unweighted mean was the preferred balance measure for cluster sizes of 10 and 25, but for the cluster size of 100, the pooled ASB as an equally weighted mean performed better (r = .19) than the unequally weighted mean (r = .13). When within-cluster matching was paired with the cluster size of 400, the pooled ASB measures were both negatively correlated with TE estimate bias (r= -.40 for the unequally weighted mean, r= -.17 for the equally weighted mean). However, this negative correlation may not be problematic because this combination of cluster size and matching method also had the lowest level of TE estimate bias.

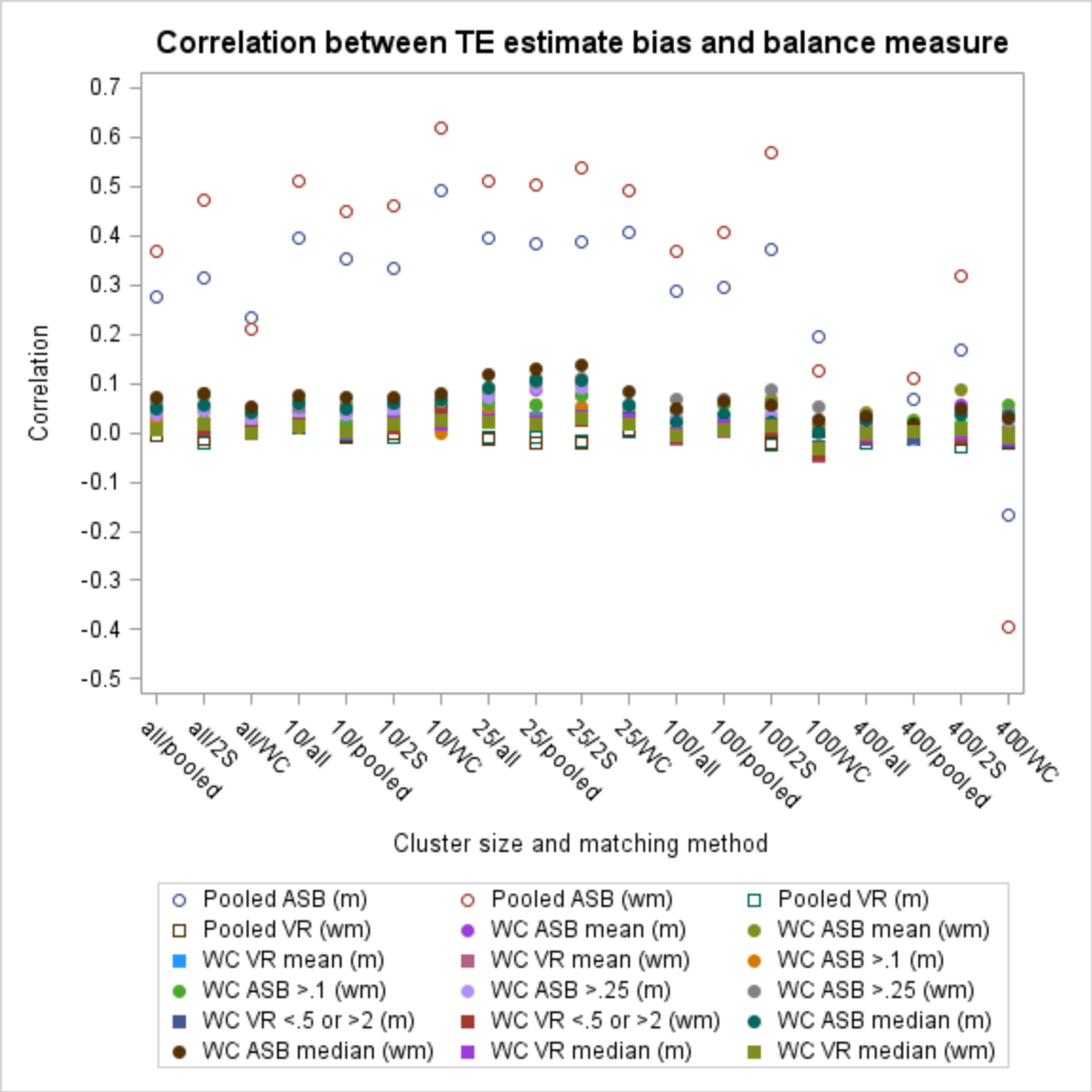

Figure 10. Correlations between treatment effect (TE) estimate bias and balance measures, by cluster size and matching method. ASB=absolute standardized bias; VR=variance ratio; WC=within-cluster balance measure; m=equally weighted mean; wm=weighted mean (according to the covariate's relation to the outcome measure). Values represent the mean correlation across ICCs and PS models.

Figure 11 illustrates the interactions between the cluster size and the PS model on the strength of the correlations between each balance measure and TE estimate bias. These results illustrate a similar pattern as in Figure 10. For most combinations of PS model and cluster size, the pooled ASB had the highest correlation with TE estimate bias, and most other balance measures had correlations close to 0. The exceptions to this were

for the conditions in which TE estimate bias was lowest. In particular, there were negative correlations between the pooled ASB and TE estimate bias when the RI model was used on a sample with 400 units per cluster (r=-.29 for the unequally weighted mean, and r=-.15 for the equally weighted mean).

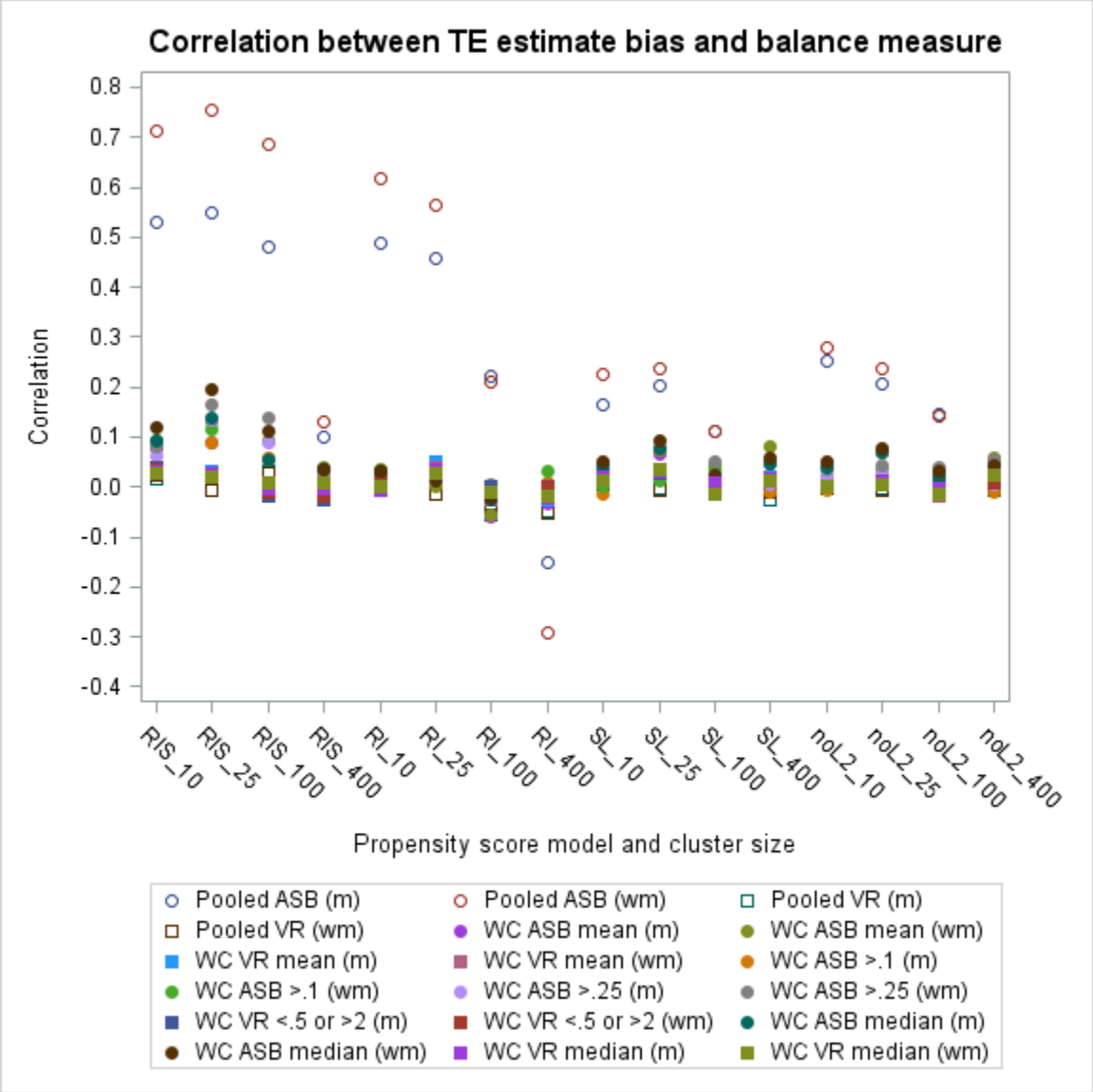

Figure 11. Correlations between treatment effect (TE) estimate bias and balance measures, by cluster size and propensity score model. RIS= random intercepts and slopes model; RI= random intercepts model; SL=single-level model; noL2= model without cluster-level covariates; ASB=absolute standardized bias; VR=variance ratio; WC=within-cluster balance measure; m=equally weighted mean; wm=weighted mean (according to the covariate's relation to the outcome measure). Values represent the mean correlation across ICCs and matching methods.

Further investigation of the negative correlations shows that there was a three-way interaction between cluster size, PS model, and matching method on the strength of the correlations between the pooled ASB and TE estimate bias. Negative correlations were observed under conditions of cluster sizes of 100 and 400 based on PS scores from some PS estimation models. Specifically, when within-cluster matching was used with a sample of 100 units per cluster, the correlation between pooled ASB and TE estimate bias was positive for propensity scores from the RIS model (r=.68) but negative for those from the RI, SL, or NoL2 models (r=-.12, -.32, -.29, respectively). When within-cluster matching was used with a sample with 400 units per cluster, the pooled ASB was negative for data from each PS estimation model, ranging from r= -.03 for the RIS model to r=-.69 for the SL model. Figure 12 shows a scatterplot in which the data from SL propensity score model were paired with within-cluster matching with the results shown grouped according to cluster size. The figure reveals that the negative correlations only occurred when there were very small levels of TE estimate bias and imbalance, which occurred with within-cluster matching and large sample sizes. Because the pooled ASB in these conditions was close to 0 across all replications, it could not be a good predictor of TE estimate bias. However, in these conditions, it is not problematic because TE estimate bias is very low.

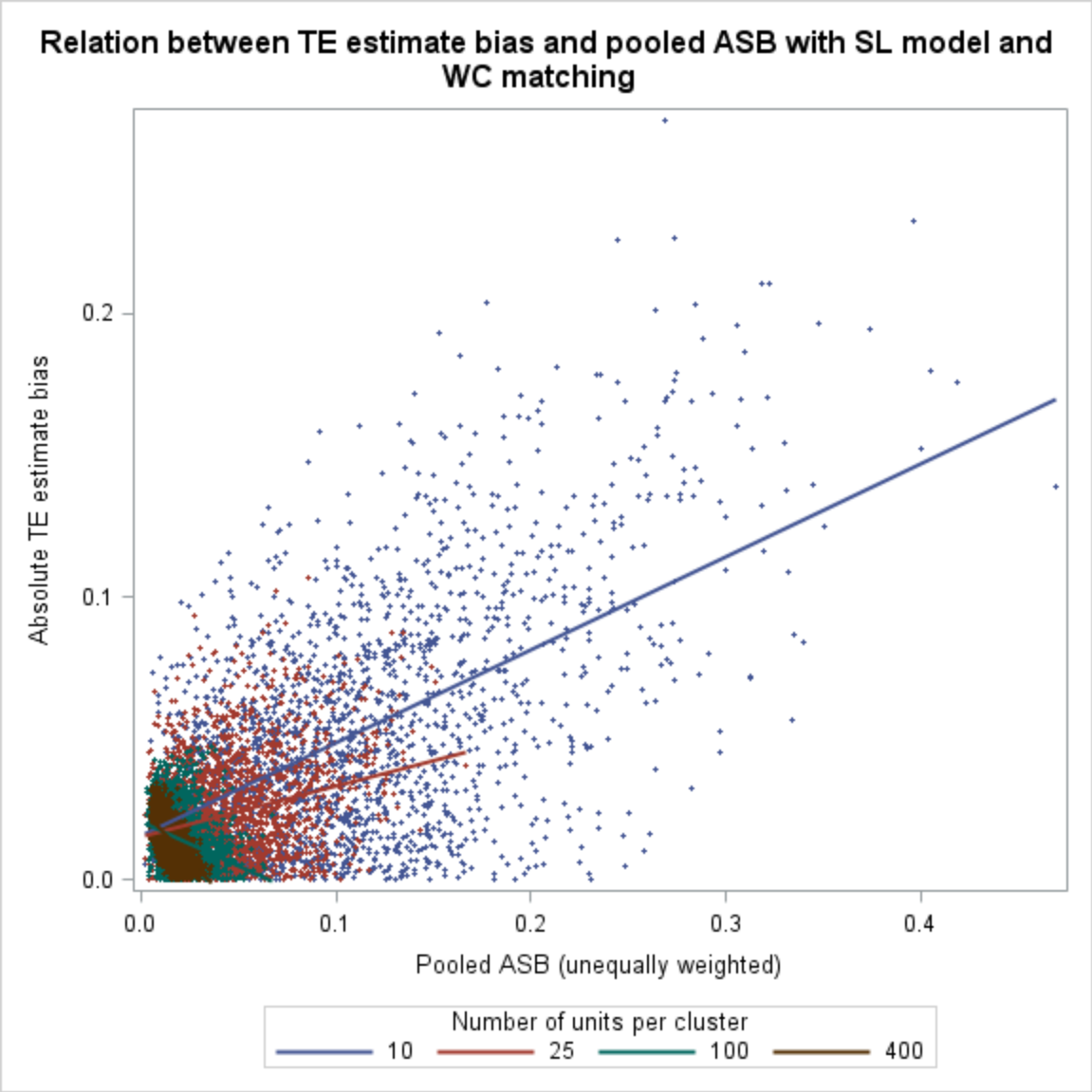

Figure 12. Scatterplot depicting the relation between absolute treatment effect (TE) estimate bias and the pooled, weighted absolute standardized bias balance measure for the conditions with the single-level (SL) propensity score model and within-cluster matching, disaggregated by cluster size.

4.5 Summary of Simulation Results

To summarize, the simulation tested which balance measures were most effective for 1) selecting the correctly specified model, and 2) correlating with absolute TE estimate bias. Across all conditions tested, the within-cluster ASB median (summarized as an unequally weighted mean across covariates) was most effective for the first

outcome, and the pooled ASB mean (summarized as an unequally weighted mean) was most effective for the second outcome. The results differed because the correctly specified PS model was not, in most cases, the model that resulted in the lowest level of TE estimate bias. Other authors have noted that the goal of PS modeling is not to achieve model fit but to achieve balanced samples, which may require researchers to use a model with relatively poorer fit than other potential PS models (e.g., Schafer & Kang, 2008). Because the goal of PS matching is to achieve unbiased TE estimates, researchers should use the results from the correlation outcome as a guide for assessing balance in multilevel samples. However, this simulation design assumes that the researcher would estimate a single TE, rather than separate TE estimates for each cluster. Estimation of heterogeneous treatment effects would likely require correct specification of potential random intercepts and slopes in a multilevel model, making the first outcome more relevant (Kim & Seltzer, 2007). This limitation and its implications are further described in Chapter 6. The next chapter provides two empirical illustrations that assume the use of homogeneous TE estimates and therefore uses the recommended balance measure from the second outcome. Researchers can use these illustrations as a blueprint for conducting multilevel PS matching, performing diagnostics, and estimating treatment effects with different cluster sizes.

Chapter 5. Empirical Illustrations

Empirical analyses were conducted to demonstrate the use of multilevel propensity score matching and balance assessment with real data and identify additional challenges with assessing multilevel PS balance that did not arise during the simulation. As described in Chapter 3, the ECLS-K:2011 (NCES, Tourangeau et al., 2015) was the basis for the parameter values used in the simulation and is one of the two empirical illustrations in this chapter. Because the multilevel PS methods and balance measures behave differently based on the cluster size, a second dataset was selected with much larger cluster sizes. Whereas the ECLS-K dataset has an average of 15 kindergarteners clustered within each school, the Health Behavior in School-Aged Children 2009-10 (HBSC) is an international survey with thousands of youth clustered within each country. The chapter provides a blueprint of the steps involved for researchers working with either small or large cluster sizes.

The chapter includes four sections: 1) an overview of the steps involved in multilevel PS matching 2) the ECLS-K illustration, 3) the HBSC illustration, and 4) a summary of results and brief conclusions. In the first section, a step-by-step flowchart (Figure 13) expands the four-step process for PS matching described in Chapter 2 into a nine-step process for multilevel PS matching with questions for applied researchers to answer as they complete their analyses. Each illustration includes an introduction to the topic and descriptions of the sample, methods, and results. The methods sections focus in particular on the diagnostic step, and the results sections demonstrate the differences in the results when researchers use the recommended diagnostic procedures based on the results in Chapter 4 and when they do not.

5.1 Overview of Steps for Multilevel PS Matching

Figure 13 illustrates the steps an applied researcher would take when conducting multilevel PS matching, beginning with step 1 in the top left box and ending with step 9 in the bottom right box. First, the researcher must identify covariates for the PS model that are likely to be related to treatment selection and the outcome, based on prior research on the topic. At this point, the researcher should also consult the literature to determine the relative importance of each covariate in predicting the outcome and assign weights to the covariates accordingly. These weights will be used in step 6. Next, the researcher should consider whether to incorporate any interactions or higher-order terms into the PS model. In step 3, the researcher should determine which combinations of multilevel PS models and matching methods to use based on the number of units in each cluster. Next, the researcher should run a set of propensity score models, including the recommended types of models in step 3 and variations of interactions or higher-order terms selected in step 2. For example, if the researcher wants to include two possible interactions and plans to use either an SL or RI model, the researcher might run eight PS models (SL or RI model paired with either no interaction, the first interaction, the second interaction, or both). Using an RIS model would also involve identifying potential random slopes based on the literature and including models with different combinations of them. In step 5, the researcher then conducts one or two matching approaches using the propensity scores obtained from each PS model. If the researcher previously ran eight PS models and subsequently uses two matching methods, this would result in 16 matched samples. Next, the researcher should determine which balance measures to use. Assuming that the researcher is estimating a single, homogeneous treatment effect, the researcher would select either the pooled ASB (for clusters that have on average less than

400 units each) or the within-cluster ASB (for clusters that have on average more than 400 units each). Next, the researcher would use the selected balance measure to calculate balance for each of the matched samples. For the within-cluster ASB, the researcher would calculate the ASB for each covariate within each cluster and then take the median across all clusters. For both the pooled ASB and the within-cluster ASB, the researcher would calculate a weighted mean across the ASBs of each covariate using the weights selected in step 1. The researcher would also examine overlap using a jitter plot, as described in Chapter 2, to ensure that all matched units in the sample do not have propensity scores outside the range of the opposite treatment selection group. If some units do not overlap with the opposite group, the researcher should return to step 5 and either add a caliper or narrow the caliper width. Finally, once the researcher has selected the most balanced sample and overlap is sufficient, the researcher can then use the selected matched sample for the TE analysis. The next two sections of the chapter will follow these steps with illustrations from the ECLS-K and HBSC datasets.

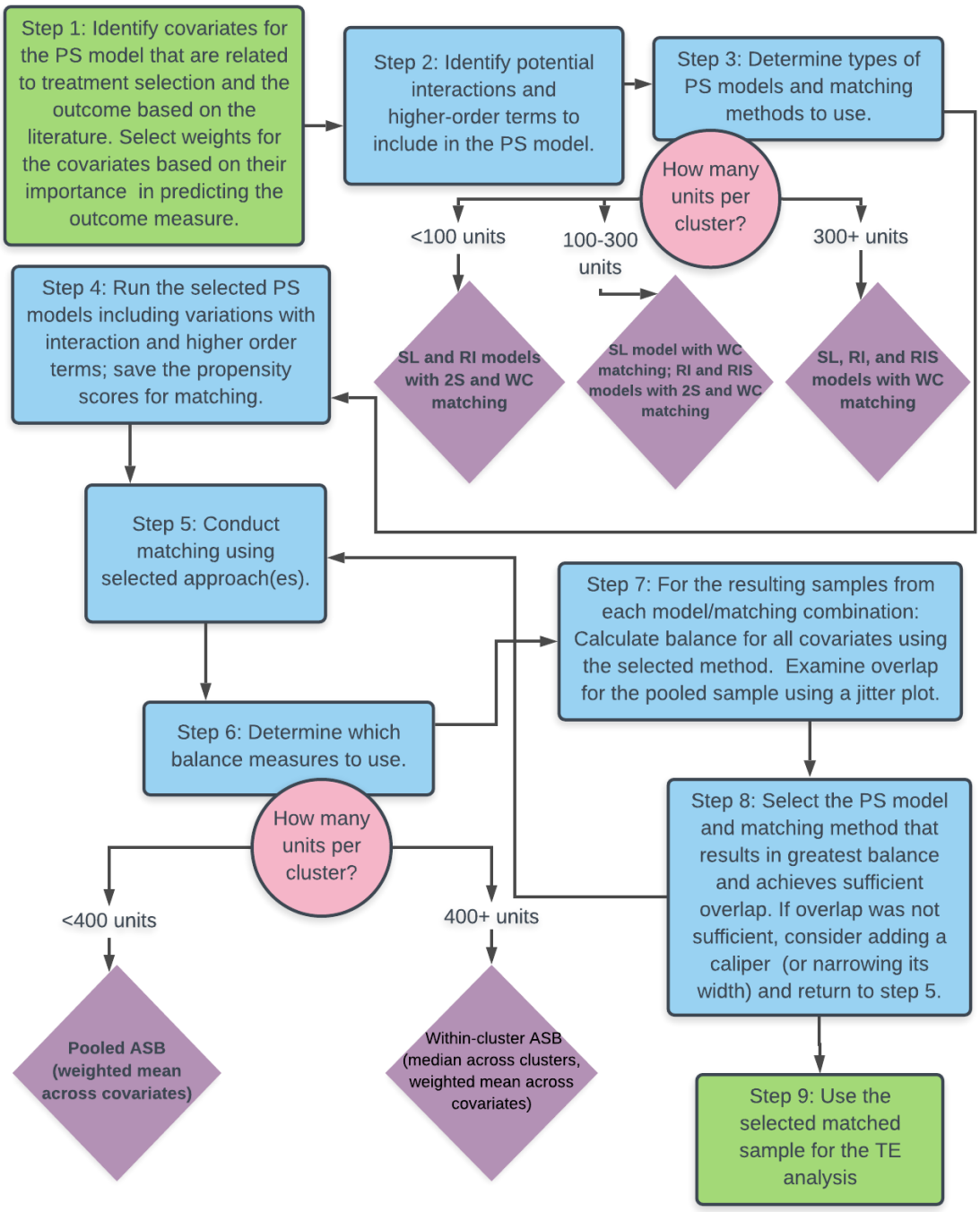

Figure 13. Flowchart illustrating the series of decisions required for multilevel propensity score matching. *Note.* PS=propensity score; SL=single-level model; RI=random intercepts model; RIS=random intercepts and slopes model; WC=within-cluster matching; 2S=two-stage matching; ASB=absolute standardized bias; TE=treatment effect.

5.2. Kindergarten Retention

Decades of research show that the policy of retaining low-performing students in the early grades leads to negative academic and social-emotional outcomes (Jimerson, 2001). Researchers studying this topic generally use quasi-experimental techniques because retention is typically a decision made based on a combination of district and school policy, the student's performance, and the judgment of the parents, teachers, principal, and other school staff. Selecting students to be retained (the treatment condition) is not typically accomplished through a random process or through the use of a variable with a clear cutoff, such as a test score. For this reason, RCT and RDD designs are not feasible, making propensity score methods optimal in this context. In a meta-analysis of 20 studies on early grade retention from the 1990s, all used matching or covariate adjustment to make causal inferences (Jimerson). Specifically, 75% of the studies matched or controlled for SES, 70% matched or controlled for gender, 65% matched or controlled for academic achievement, and 30% matched or controlled for social-emotional factors. Later, Wu, West, and Hughes (2010) used propensity score matching based on 72 baseline characteristics to understand the effects of first grade retention on short- and long-term social-emotional outcomes. Hong and colleagues expanded upon this research by using the large, nationally representative ECLS-K 1998 sample to assess the effect of kindergarten retention on a variety of later academic and social-emotional outcomes using multilevel PS stratification (Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007, 2008). The present analysis uses multilevel PS matching to assess the effects of kindergarten retention on reading outcomes. Although this analysis is unlikely to produce novel findings that will result in policy changes, it is intended only to illustrate how the techniques examined in the simulation can be applied to a context that

lends itself to multilevel PS methods, which will likely be useful for researchers working with this and similar NCES datasets or administrative data.

5.2.1 Data. Whereas Hong and colleagues (Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007, 2008) used the ECLS-K 1998 cohort, this analysis used the most recent cohort of children who began kindergarten in 2011. The analytic sample was limited to children who attended a school that had a policy that allowed for kindergarten retention and did not have a missing test score in the reading achievement test at the end of the treatment year. The sample used for matching included 857 schools and 12,451 students (an average of 15 students per school).

5.2.2 Variable selection. The PS models included a comprehensive set of variables from the ECLS-K dataset that were likely to predict retention and/or reading achievement based on Hong and Raudenbush (2005). Their analysis used more than 200 variables, including student/parent, classroom, and school characteristics, but for the purpose of illustration, this analysis was limited to a two-level model with students at the unit level and schools at the cluster level. Student characteristics included gender, ethnicity, age at kindergarten entry, socio-economic status (SES), type of child care received prior to kindergarten, number of parents in the household, number of siblings in the household, disability status, kindergarten teacher's perception that the child fell behind due to frequent absences, participation in extracurricular activities, parent's and teacher's ratings of educational expectations of the child, availability of a home computer, availability of children's books in the home, literacy and math knowledge at the beginning of kindergarten, parent and teacher ratings of social skills, parent and teacher ratings on emotional and behavioral problems, and participation in pull-out

instruction. The school characteristics included school size, school type (public or private), percentage of Hispanic children, attendance rates, availability of services for children with disabilities, adequacy of facilities rating, and a school safety rating.

5.2.3 Propensity score models and matching procedures. Using these variables, several multilevel PS models were constructed and compared for balance. The analysis limited the methods to those that had reduced bias most effectively for small cluster sizes based on the results of the simulation and other simulation studies (Rickles & Seltzer, 2014; Arpino & Cannas 2016). This included the SL model that included school-level covariates and the RI model (Figure 13, step 3); the RIS model was not tested due to its poor performance with clusters of this size. The models tested also varied in whether they included interactions between kindergarten reading and kindergarten math scores and between kindergarten reading and age at kindergarten entry. In all models, student-level variables were centered at the school mean, and school-level variables were centered at the grand mean. The school-level averages of the student-level variables were not included in the model.

Two-stage and within-cluster matching were both conducted using the propensity scores from each of the PS models. Because there were not any conditions in the simulation for which pooled matching was most effective for reducing TE estimate bias, pooled matching was not considered. All matching methods utilized nearest neighbor matching without replacement and a caliper of .2 standard deviations of a propensity score.

In total, 16 variations of models and matching methods were conducted: two model types (SL or RI) by four combinations of interactions (one model with no

interactions, two models with one interaction each, and one model with both interactions) by two matching methods (two-stage and within-cluster).

5.2.4 Diagnostics. In the diagnostic step, the balance for all combinations of the PS models and matching methods were calculated using both the recommended and a few of the non-recommended balance measures from the simulation. By assessing balance with the non-recommended measures, one can determine whether using a non-recommended balance measure would lead to selecting a different matched sample and therefore obtaining a different TE estimate. Based on the results from the correlations between balance measures and TE estimate bias from the simulation, the recommended balance measure was the pooled ASB, with covariates summarized as a mean weighted by the likely importance of the covariates in predicting the outcome (Figure 13, step 6). For the purpose of this review, weights were assigned based on the importance of the covariate in the WWC Review Protocol for Beginning Reading Interventions, Version 3.0 (U.S. Department of Education, 2014). The protocol prioritizes baseline equivalence for a pre-intervention measure and secondarily considers gender, race, English learner status, disability status, SES, location type (urban, rural, suburban), and average class size (small, medium, large). Weights were assigned such that the kindergarten reading achievement variable was equal to the total weight of the five secondary variables from the WWC protocol, which was equal to the total weight of the 31 covariates that did not correspond to a variable in the protocol. The non-recommended balance measures included the equally-weighted pooled ASB, pooled VR, the pooled ASB >.1 indicator, the pooled ASB >.25 indicator, the pooled VR > 2 or <.5 indicator. For the non-recommended balance measures, the balance across covariates was summarized using an

equally weighted mean. For this empirical example, within-cluster balance measures could not be calculated because there were too many instances in which only one or two students from a school were included in the matched sample; these measures are included in the second empirical example presented in Section 5.3.

Table 9

Assessment of balance of PS model and matching combinations used to select the sample for the ECLS-K analysis of the effects of kindergarten retention on reading outcomes

PS model and matching method				Balance measures					
Matching method	Model type	Model interactions	N	ASB (w)	ASB	VR distance	ASB > .1	ASB > .25	VR < .5 or > 2
Two-stage	SL	None	860	.043	**.033**	.073	**.033**	**.003**	**.000**
		Reading*math	852	.053	.043	.083	**.033**	**.003**	**.000**
		Reading*age	852	.053	.043	.083	.083	**.003**	**.000**
		Both	848	.043	.043	.093	**.033**	**.003**	**.000**
	RI	None	674	.053	.043	.103	**.033**	**.003**	**.000**
		Reading*math	674	.053	.053	.113	.143	**.003**	**.000**
		Reading*age	678	**.033**	.043	.093	.053	**.003**	**.000**
		Both	668	.043	.053	.103	.113	**.003**	**.000**
Within cluster	SL	None	320	.073	.053	.083	.163	**.003**	**.000**
		Reading*math	324	.073	.053	.073	.163	**.003**	**.000**
		Reading*age	334	.073	.053	.083	.163	**.003**	**.000**
		Both	318	.043	.043	.063	.053	**.003**	**.000**
	RI	None	246	.083	.083	.073	.353	.083	**.000**
		Reading*math	248	.053	.063	**.063**	.193	**.003**	**.000**
		Reading*age	260	.063	.053	.093	.193	**.003**	**.000**
		Both	250	.053	.043	.083	.053	**.003**	**.000**

Note. PS=propensity score. N= number of students in the matched sample. SL=single-level model (includes cluster-level covariates). RI= random intercept model. ASB=absolute standardized bias. ASB (w) = absolute standardized bias, weighted according to importance of covariates in predicting the value of the outcome. VR=variance ratio.

The bold represents the most balanced sample according to each balance measure. Multiple are in bold if there was an exact tie between samples.

The variance ratio was calculated as the distance of the ratio from 1 such that lower values indicate greater balance.

Proceeding to the diagnostic step (Figure 13, step 7), each of the PS models and matching methods were compared for balance using the pooled ASB with covariates weighted according to their importance to select the sample. As shown in Table 9, the RI

propensity score model that includes the reading and age interaction paired with two-stage matching had the lowest level of imbalance, and was thus selected. The table also shows that using a different balance measure would lead to selecting different matched samples. Using the pooled ASB with covariates assigned equal weights would lead the researcher to choosing the sample derived from the SL model without any interactions with two-stage matching. Finally, using the pooled VR would lead one to selecting the RI model with the reading and math interaction and within-cluster matching.

Consistent with the simulation results, the indicators of ASB >.1, ASB >.25 and VR < .5 or >2, which were calculated as the proportion of covariates that met the threshold, were not effective for selecting a single best model. The ASB > .1 resulted in little differentiation between models and a four-way tie between PS methods. The remaining indicators performed much worse, as nearly all combinations of models and matching methods had a value of 0.

Additionally, overlap was examined through the visual display of propensity scores in a jitter plot for the pooled sample. Figure 14 shows the jitter plot of the pooled sample from the RI propensity score model with the reading and age interaction and two-stage matching, which was selected by the recommended balance measure (the pooled ASB, summarized across covariates as an unequally weighted mean). The jitter plot shows that retained students were matched with promoted students with propensity scores in a similar range. However, a high percentage of retained students were excluded from the sample because no promoted students had similar propensity scores. This is problematic; the researcher no longer can claim that the treatment effect is the ATT, since it only applies to a smaller subset of those treated. However, using a less restrictive

matching method that would include these retained students would lead to greater levels of imbalance and therefore biased TE estimates. An analyst deciding to continue with using the matched sample in Figure 14 should caution readers that the TE estimate only applied to a subsample of the treated and was therefore not a true ATT.

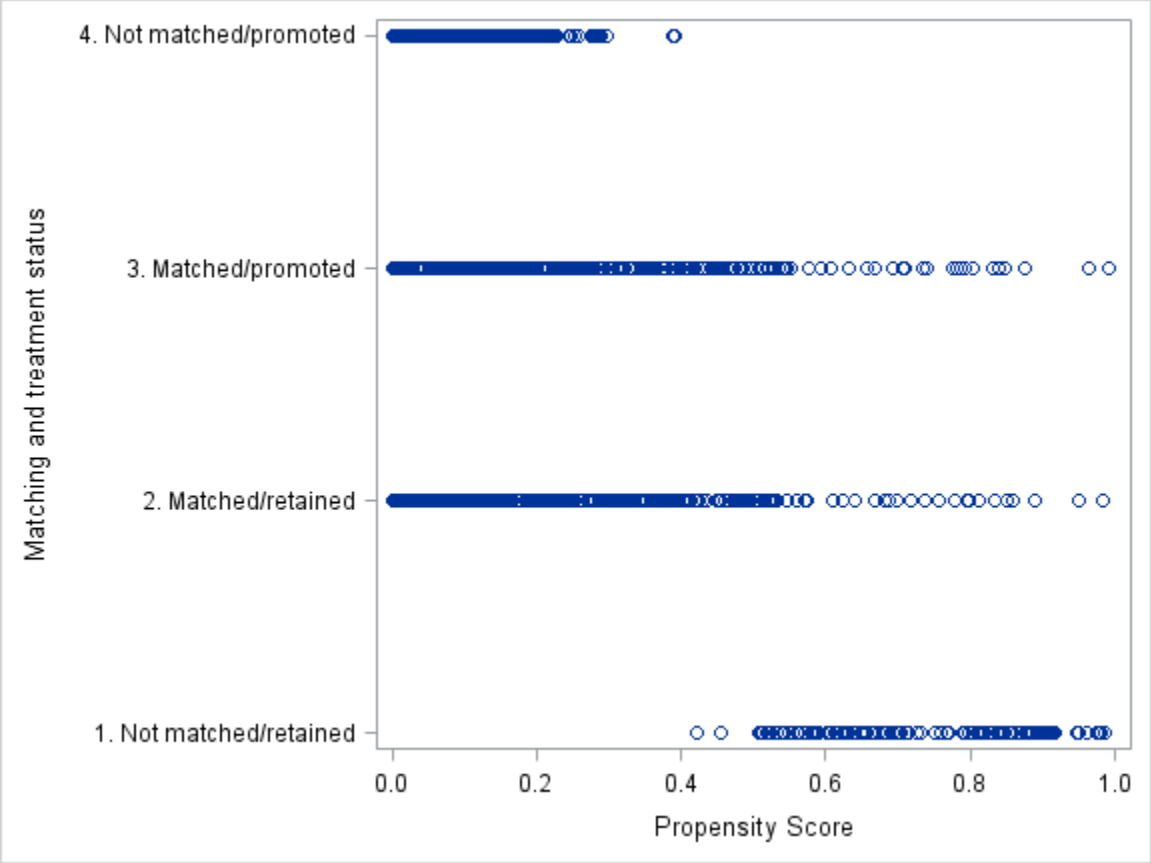

Figure 14. Overlap of matched sample from the random intercepts model with the reading and age interaction and two-stage matching.

5.1.5 Results. Finally, the TE was estimated in two ways: (1) by comparing means between retained and promoted students and (2) by using a regression model that controlled for baseline covariates. Assuming that the matched sample is balanced, the difference in treatment group means should be sufficient. However, controlling for covariate differences in both the matching phase and the TE estimation phase can lead to more accurate results if there are misspecifications in one of the models (Funk et al.,

2011). By using both methods, one could examine the extent that controlling for baseline covariates in the TE model would affect the final TE estimate. The TE was estimated separately for the unmatched sample and three matched samples resulting from using different balance measures: 1) RI model that included the reading and age interaction and used two-staged matching (the most balanced sample based on the pooled ASB with covariates weighted according to their importance), 2) SL model without interactions that used two-stage matching (the most balanced sample based on the pooled ASB with covariates weighted equally), and 3) RI model with the reading and math interaction and within-cluster matching (the most balanced sample based on the pooled VR with covariates weighted equally). The first matched sample represents the one selected from the recommended balance measure and the second and third matched samples represent the samples from alternative, non-recommended balance measures.

As shown in Table 10, the choice of methods made a large impact on the TE estimate based on the simple comparison of means. Without any matching or covariate adjustments, the TE estimate indicated that retained children did worse in first grade reading than promoted children by more than 1.5 standard deviations. The standardized TE estimate was reduced to -.38 with the use of the preferred PS model and matching method based on the pooled ASB (with covariates unequally weighted). The standardized TE estimate from the three matched samples without the use of covariate adjustment ranged from -.38 to -.57 for the sample with within-cluster matching and the RI model that included the reading and math interaction.

The differences between TE estimates across the PS methods were much smaller with the use of a TE regression model that controlled for baseline covariates. In this case,

the difference in effect size between the full, unmatched sample and the sample from the recommended strategy differed by less than .01 standard deviations of a first grade reading score. With the use of regression models to estimate the treatment effects, the difference in standardized TE estimates between the matched samples that used two-stage matching was also much smaller (.03). However, the standardized TE estimate from the sample that used within-cluster matching (-.60) was noticeably stronger than the other TE estimates, as it fell outside of the confidence intervals of two out of the other three TE estimates. Because the true TE estimate is unknown in the case of an empirical analysis, this causes one to wonder if certain PS models and matching methods that heavily trim the sample lead to greater levels of bias and imbalance than simply controlling for the variation using regression adjustment alone. As shown in Table 10, the sample that used within-cluster matching was much smaller than the others, and the resulting estimate should be considered a subset of the ATT rather than a true ATT.

Table 10

Difference in standardized reading scores between those retained and those not retained in kindergarten

PS model and matching method					Effect size (D) and 95% CI	
Match	Model	Interactions	Balance measure used	N	Estimate from comparison of means	Estimate from regression model with covariates
None	NA	NA	NA	12,451	-1.66 (-1.75, -1.56)	-.48 (-.57, -.39)
2S	RI	Read*age	Pooled ASB(w)	678	-.38 (-.53, -.23)	-.48 (-.63, -.33)
	SL	None	Pooled ASB	860	-.49 (-.62, -.35)	-.45 (-.59, -.32)
WC	RI	Read*math	Pooled VR	248	-.57 (-.82, -.31)	-.60 (-.78, -.42)

Note. 2S=two-stage matching. WC=within-cluster matching. SL=single level. RI=random intercept model. NA=not applicable. ASB=absolute standardized bias. ASB(w) = absolute standardized bias, weighted according to importance of covariates in predicting the value of the outcome. VR=variance ratio. CI=confidence interval.

Using the recommended balance measure to select the sample and then estimate the TE using a regression model with covariate adjustment would lead to the conclusion that being retained in kindergarten would lead to lower reading achievement by .48 standard deviations. However, because of the wide range of estimates across the PS models and matching methods, another approach would be to calculate a model average estimate, rather than relying solely on one estimate. With this approach, one could weight each TE estimate according to the likelihood of it being correct. This alternative approach will be further discussed in the next chapter.

5.3 Bullying Victimization

The next empirical illustration used the HBSC dataset to examine the effects of bullying victimization on life satisfaction. This analysis was included because it could demonstrate the use of multilevel PS matching with much larger clusters (thousands of youth within countries) in comparison to the first analysis. The survey is administered to youth in 40 countries and/or regions and covers topics related to physical and behavioral health, education, social and sexual behavior, and alcohol and drug use. Several studies have examined trends in bullying victimization using the HBSC dataset. For example, Lian et al. (2018) examined the relations between being bullied and weight status and body self-image using logistic regressions. The authors made covariate adjustments for SES, family structure, and classmate support and conducted separate analyses for males and females. Another study used HBSC data to compare rates of bullying and victimization across 40 countries and examine cross-national trends by sex and age group (Craig et al., 2009). These studies both used regression techniques rather than PS methods, but studies in other topic areas have used HBSC data with PS methods to make

causal inferences (e.g., Elstad & Pedersen, 2012; Winter, Combs, & Ward, 2018), and it is likely that applied researchers will continue to do so.

5.3.1 Data. The analytic sample was limited to youth who had non-missing data for the outcome variable and the selected covariates, which included 29 countries and 104,181 youth.

5.3.2 Variable selection. The PS models included a set of variables from the HBSC dataset that were likely to predict being a victim of bullying and ratings of life satisfaction based on prior research (Craig et al., 2009; Lian et al., 2018). This included youth characteristics, including gender, age, grade level, overweight status (based on reported body mass index [BMI]), physical activity, SES, a composite measure of classmate support, and another composite measure of peer support. SES and classmate support were constructed using the same items as in Lian et al., and the peer support measure was constructed using the set of items in the peer support scale. To incorporate relevant country-level covariates, the HBSC data were merged with data from the World Bank (available at data.worldbank.org) on each country's per capita gross domestic product (GDP) and the Gini coefficient, a measure of a country's level of wealth inequality. Interactions between age and grade and between age and gender were also examined.

5.3.3 Propensity score models and matching procedures. Using these variables, several PS models and matching methods were compared for balance. The analysis limited the methods to those that were most effective for reducing TE estimate bias with large cluster sizes. Specifically, variations of SL, RI, and RIS models were each paired with within-cluster matching (Figure 13, step 3). Because even the smallest

country included nearly 1,000 youth, there was no reason to match youth across different countries. Moreover, when comparing countries of youth who have a wide range of social and cultural norms, it is likely that they may have different understandings of a concept like bullying, which would make controlling for all country-level differences through modeling difficult. In this context, within-cluster matching can be more effective for controlling for the cultural differences than controlling for differences with cluster-level covariates. The RIS models included random slopes for gender, SES, and age. If there was an interaction in the model, these were also included as random slopes. All youth-level variables were clustered at the country mean, and country-level variables were clustered at the grand mean.

In total, there were 12 variations of models and matching methods tested: three model types (SL, RI, RIS) by four combinations of interactions (one model with no interactions, two models with one interaction each, and one model with both interactions) by one matching method (within-cluster).

5.3.4 Diagnostics. Based on the results from the simulation, the recommended balance measure was the within-cluster ASB, summarized as a median across clusters and as a weighted mean across covariates. To calculate this measure, the ASB of each covariate is calculated within each cluster; next, the median ASB across the clusters is calculated for each covariate; and finally, the median ASB for all covariates are aggregated as a weighted mean according to each covariate's likely importance in predicting the outcome. Peer support, SES, gender, and age were given a weight equal to twice of the remaining covariates due to their importance in this literature (e.g., Craig et al., 2009). The non-recommended balance measures included the pooled ASB and VR,

the within-cluster mean VR, and the within-cluster indicator of VR > 2 or <.5. The within-cluster mean ASB and the within-cluster ASB indicator of >.1 performed similarly to the ASB median in large sample sizes, so these balance measures were compared to determine whether any of them would result in selecting a different PS model and matching method. For each of the non-recommended measures, the balance across covariates was summarized using an equally weighted mean. The pooled ASB indicators of >.1 and >.25 and the pooled VR indicator of <.5 or >2 were not calculated, because based on the simulation results, it was unlikely that any covariates would have values outside these thresholds across the pooled sample.

Table 11

Assessment of balance of PS models used to select the sample for the HBSC analysis of the effects of kindergarten retention on reading outcomes

PS model			Within-cluster balance					Pooled balance	
Model type	Interactions	N	Median ASB (w)	Mean ASB	ASB > .1	Mean VR distance	VR >2 or <.5	ASB	VR distance
SL	none	49,384	.054	.065	.198	.062	**.000**	.005	.020
	age*grade	49,264	.058	.067	.202	.058	**.000**	.004	.011
	age*gender	49,384	.057	.065	.210	.062	**.000**	.005	.020
	both	49,286	.054	.061	.202	**.054**	**.000**	.004	.010
RI	none	49,454	.058	.066	.214	.065	**.000**	.005	.021
	age*grade	49,366	.055	.064	.206	.061	**.000**	.004	**.009**
	age*gender	49,458	.056	.068	.226	.066	**.000**	.006	.022
	both	49,374	.056	.065	.206	.060	.004	.005	.012
RIS	none	49,176	**.036**	**.043**	**.056**	.066	**.000**	.005	.022
	age*grade	49,234	.049	.060	.103	.057	**.000**	.005	.012
	age*gender	49,376	.042	.052	.155	.064	**.000**	**.004**	.022
	both	49,214	.044	.053	.147	.054	**.000**	.007	.013

Note. PS=propensity score. N= number of students in the matched sample. SL=single-level model (includes cluster-level covariates). RI= random intercepts model. RIS= random intercepts and slopes model. WC=within-cluster balance measure. ASB=absolute standardized bias. ASB (w) = absolute standardized bias, weighted according to importance of covariates. VR=variance ratio.

All matching was carried out within clusters.

As shown in Table 11, based on the recommended unequally-weighted, within-cluster ASB median, the sample derived from the RIS model without the interactions should be selected for estimating the treatment effect. Using either the within-cluster ASB mean or the within-cluster indicator of ASB > .1, would lead to the same conclusion. However, using one of the non-recommended balance measures would lead to selecting samples derived from other PS models. Using the within-cluster VR would lead to selecting a single-level model that included both interactions, the pooled ASB would result in selecting the RIS model with the age and gender interaction, and the pooled VR would lead to selecting the RI model with the age and grade interaction. Because clusters were very large, there were few covariates that had VRs below .5 or above 2 for any of the clusters and models. For this reason, the within-cluster indicator of VR <.5 or >2 could not identify a preferred model. Similarly, the pooled ASB was close to 0 across the pooled sample for each model, ranging from .004 to .007.

Because within-cluster balance measures were best suited for this analysis, it also seemed appropriate to review overlap within clusters, rather than across the pooled sample. These initially were reviewed as separate figures for each country, but were summarized into two plots (Figure 15). The top plot shows the overlap in propensity scores among youth who were matched, and the plot beneath it shows the overlap in propensity scores among youth who were not matched. Both plots are disaggregated by country. As shown in the figure, the matched sample of those who were bullied and those who were not is sufficiently overlapping, whereas the unmatched sample is clearly separated with those who were not bullied with propensity scores lower than nearly all propensity scores of those who were bullied. Because thousands of youth who said they

were bullied remained unmatched, this has implications for interpreting the ATT. As with the matched sample from ECLS-K illustration, this matched sample should be treated as a subset of the ATT. One can also notice differences in the distribution of propensity scores and the prevalence of bullying by country, which reinforces the decision to match within countries. For example, no one from Turkey was included in the matched sample because none reported that they had been bullied. The different distributions in propensity scores may reflect differences in how the concept of bullying translates across different languages and cultures.

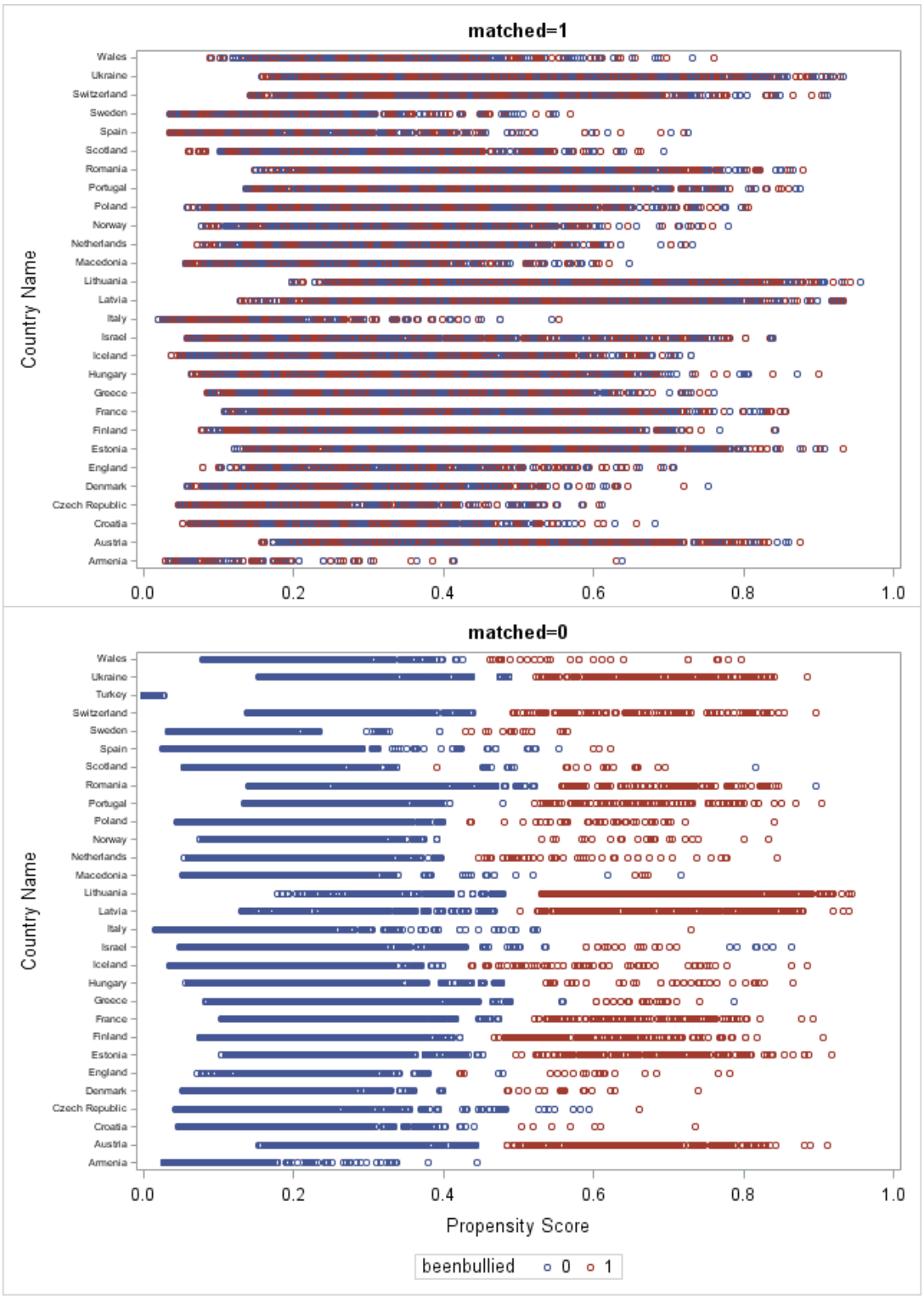

Figure 15. Overlap in propensity scores by country, matching status, and treatment status (bullied or not). Overlap is shown for the selected sample with the RIS propensity score model with no interactions.

5.2.5 Results. Treatment effects were estimated in the same two ways as with the ECLS-K sample—as a difference in means and as a regression model that controlled for baseline differences. Table 12 presents the standardized effect size estimates from each outcome analysis.

Table 12

Difference in ratings of life satisfaction between those who had been bullied and those who had not been bullied

Model type	Interaction	Balance measure used to select the sample	Sample size	Estimate from comparison of means		Estimate from regression model with covariates	
				Effect size (D)	95% CI	Effect size (D)	95% CI
None	NA	NA	104,184	-.348	(-.362, -.334)	-.251	(-.265, -.237)
RIS	None	WC ASB median	49,176	-.269	(-.287, -.252)	-.269	(-.287, -.251)
	Age*gender	Pooled ASB	49,234	-.258	(-.276, -.241)	-.260	(-.278, -.243)
SL	Both	WC VR	49,286	-.265	(-.283, -.248)	-.264	(-.282, -.247)
RI	Age*grade	Pooled VR	49,366	-.264	(-.282, -.246)	-.263	(-.280, -.245)

Note. PS=propensity score. SL=single-level model (includes cluster-level covariates). RI= random intercepts model. RIS=random intercepts and slopes model. CI=confidence interval for the effect size.

All matching was conducted within clusters.

As expected, the standardized difference in means on a scale of life satisfaction was greatest in magnitude without the use of covariate adjustment or matching (-.348). However, the difference between the unmatched sample and the matched samples were much smaller compared with the ECLS-K retention analysis, suggesting that there was less selection bias compared with ECLS-K. Across the matched samples, there were only small differences between the TE estimates when calculating as either a difference in treatment and control means or using a regression model with covariates. When both matching and covariate adjustment to the TE model were employed, the range in standardized effect sizes was less than .01. One might conclude that because of the large

within-cluster sample sizes and the use of within-cluster matching, all of the matched samples were well-balanced, and therefore, small changes to the propensity score model were unlikely to make a difference in the TE estimate. For this reason, using the correct balance measure was less critical than was the case in the ECLS-K illustration.

5.4 Summary of Empirical Analyses

This chapter presented two different types of multilevel PS matching analyses, which both followed the process outlined in Section 5.1 and Figure 13. In both illustrations, a series of different types of PS models and matching methods were assessed for balance. Based on the results of the simulation and the average cluster size in the dataset, the recommended balance measure was used to select the sample for the TE analysis. Then the TE was estimated using the sample from the recommended balance measure, and the effects were compared to those when a different balanced sample was selected or when matching was not used at all.

In the ECLS-K analysis, using different balance measures led to different selections of which PS method to use and subsequently a wide range of TE estimates. Even when controlling for baseline covariates in the TE model, the standardized difference in treatment effects ranged from -.45 to -.60 with the three matched samples. In the HBSC analysis, the difference in effects between using a recommended balance measure and another balance measure to select the sample was much smaller (no more than .01 standard deviations when combined with regression adjustment). In the context of the ECLS-K analysis, the balance measure choice and the resulting matched sample based on that choice had a greater impact on the results, presumably because of the small cluster sizes and the greater degree of selection bias. The small cluster sizes and low

prevalence of retention within each cluster made within-cluster matching and multilevel modeling difficult.

In both analyses, the unmatched sample resulted in a similar TE estimate as the recommended matched sample when the same covariates were used in the regression-based TE model. This begs the question of why one would use PS matching if a researcher would likely obtain the same results from matching as with simply including all of the covariates in the TE model. It is possible that in both analyses the selected covariates made treatment selection strongly ignorable; therefore, covariate adjustment alone was sufficient for reducing TE estimate bias. In other situations, where there are more unmeasured confounders, it is likely that using PS matching to trim the sample would be more effective for reducing TE estimate bias.

Chapter 6. Discussion

Multilevel settings present a unique challenge for those wishing to conduct propensity score (PS) matching. When treatment is assigned at the unit level but the rates and predictors of treatment selection differ across clusters, researchers must account for the nested structure of the data through the PS model, the matching process, or both. Prior methodological studies have proposed and tested approaches for modeling the propensity score and conditioning on the propensity score through matching or stratification (e.g., Arpino & Cannas, 2016; Arpino & Maelli, 2011; Rickles & Seltzer, 2014; Thoemmes & West, 2011). In the modeling phase, researchers can account for clustering through use of a single-level model that includes cluster-level covariates, a fixed effects model, a random intercepts (RI) model, or a random intercepts and slopes (RIS) model. In the matching phase, researchers can account for clustering through within-cluster matching (treated units can only be matched to control units within the same cluster) or two-stage matching in which matches are first attempted within the same cluster but if no adequate matches exist, a second attempt is made to match in the pooled sample. The recommended procedure depends on the extent to which the treatment selection probabilities vary across clusters, the extent to which the strength of the predictors of treatment selection vary across clusters, and the number of units per cluster.

Prior to this study, methodological researchers had not yet examined an essential component of multilevel PS matching—diagnostics for evaluating the quality of the matched sample. The diagnostic step is an essential component of PS matching because it provides evidence that the treatment effects are estimated without bias. During the diagnostic step, the researcher must evaluate balance for each potential PS model and

matching method to select a sample for analysis. However, there are many different methods for evaluating balance in a multilevel setting, and each method may result in a different conclusion about which sample to select for TE estimation. For example, researchers must choose whether to use the absolute standardized bias (ASB) or the variance ratio (VR), and whether to calculate the balance across the pooled sample or calculate it separately for each cluster. This study expanded the literature on multilevel PS matching by assessing the extent to which each balance measure could select the correctly specified PS model and the extent to which the balance measure correlated with TE estimate bias using a Monte Carlo simulation. It also demonstrated the use of the recommended and non-recommended balance measures with two empirical datasets. The purpose of this chapter is to summarize the findings from the simulation and the empirical illustrations and to discuss limitations of the study, implications, and directions for future research.

6.1 Summary of Key Findings

Across all of the manipulated factors of the simulation, one balance measure emerged as optimal for each of the assessed outcomes. For the outcome of selecting the correctly specified model, the within-cluster ASB median performed best, and for the outcome of correlation with TE estimate bias, the pooled ASB performed best. For both outcomes, the mean ASB across covariates weighted according to the strength of the covariate's relation to the outcome performed better than the equally weighted mean. There were some notable differences according to the conditions tested, especially related to cluster size. For example, the pooled ASB negatively correlated with TE estimate bias when clusters had more than 400 units each; in this case, the within-cluster ASB was

more effective for predicting TE estimate bias. However, even then, the balance measure that had the strongest correlation with TE estimate bias had a correlation near 0. In the case of the large cluster sizes, the resulting TE estimate bias was very low across the PS models and methods tested, making the choice of the correct model less critical.

The differences between the two simulation outcomes occurred because, for most conditions of the study, the correctly specified PS model was not the same as the model that led to the greatest reduction in TE estimate bias. The correctly specified model was an RIS model, the model used to generate the propensity scores for all of the simulation conditions. However, the simulation showed that the PS model that resulted in the lowest TE estimate bias depended on the cluster size. On average, the SL model resulted in the lowest TE estimate bias with clusters of 10 units each, the RI model resulted in the lowest TE estimate bias with clusters with 25 or 100 units each, and the RIS model resulted in the lowest TE estimate bias with clusters of 400 units each. Similarly, Arpino and Cannas (2016) also observed that using an RIS propensity score model led to greater levels of TE estimate bias compared to fixed effects and SL models, except when there were more than 300 units per cluster. It is not yet clear what mechanism causes the RIS and RI propensity score models to have higher levels of TE estimate bias than the SL model with small cluster sizes even when the multilevel models had better fit. Prior research on the cluster sizes required for multilevel modeling suggests that random intercept and slope parameters can be estimated with low levels of bias for samples with small cluster sizes, including some clusters with only one unit each (Bell, et al., 2010). More research is needed to understand the reasons for the cluster size requirements when multilevel PS models are used for matching.

As expected, when the correctly specified model was also the model that most reduced TE estimate bias, the recommended balance measures for the two outcomes were similar. For example, with 400 units per cluster, TE estimate bias was lowest for the correctly specified (RIS) PS model. In this case, the within-cluster ASB was the balance measure that was both the best for selecting the correctly specified model and the one that had the strongest, positive correlation with TE estimate bias.

The results from the second outcome, the correlation between the balance measure and TE estimate bias, were the basis for recommendations for researchers and for the empirical illustrations in Chapter 5. Because the goal of PS matching is to reduce imbalances and therefore TE estimate bias, researchers should use the PS model that accomplishes that, even if it means selecting a model that does not have the best relative fit with the data. Therefore, the empirical illustrations demonstrated the use of the balance measures that achieved the highest correlations with TE estimate bias in the simulation. Because the results from the simulation differed according to cluster size, a different balance measure was preferred for each analysis given their disparate cluster sizes. For the ECLS-K dataset, which included an average of 15 students per school, the preferred balance measure was the pooled ASB. By contrast, for the HBSC dataset, the preferred balance measure was the within-cluster ASB because there was an average of 3,592 youth per country. The assumption was that the HBSC dataset was most similar to the results for the simulation condition of 400 units per cluster, whereas the ECLS-K dataset was most similar to the results from the simulation conditions of 10 and 25 units per cluster.

6.2 Limitations, Implications, and Future Directions

Noting the limitations of the study is not only useful for interpreting the results, but also for generating ideas for future research that would fill gaps in the current literature. First, the study did not test all possible diagnostics or propensity score methods that could be used with assessing multilevel data. For example, the simulation did not assess overlap measures or visual diagnostics, which are both important in assessing SL propensity score models (e.g., Stuart, 2010). The empirical illustration proposed ways of incorporating visual overlap diagnostics into multilevel PS matching, but more research is needed to assess the application of overlap measures to multilevel settings.

Additionally, the simulation focused on the use of a logistic PS model with nearest-neighbor matching, because it is the most common PS method in both single-level and multilevel studies (Thoemmes & Kim, 2011). This means that the results are applicable to a large percentage of researchers who employ these methods; however, it is not clear how well the results translate to other PS models and conditioning approaches. The study used nearest neighbor 1:1 matching without replacement with a caliper of .2 standard deviations of a propensity score, so it is unknown whether results would apply to other matching procedures such as matching with replacement, *k*:1 matching, matching without a caliper, and optimal matching. In addition, all PS models used logistic regression, although probit regression and boosted modeling (Lee et al., 2010) are alternative ways of producing PS estimates. Although the results are most likely relevant to stratification, weighting, covariate adjustment, and other types of matching, more research would be needed using these methods in order to make recommendations for diagnostics to use with them.

The parameters for generating the propensity scores and outcomes in the simulation were generally based on estimates obtained from models imposed on the ECLS-K data. However, the ratio of the treated to untreated students was altered to increase the number of treated students within a cluster. This decision allowed for a greater proportion of the sample in the simulation to be matched within the pooled sample and within clusters. Although within-cluster balance measures could be assessed within the simulation, this was not the case for the empirical illustration that used the same dataset. For the ECLS-K illustration, typically between zero and two children were retained in any given school, which meant that calculating the VR or ASB within schools was not an option. Similarly, Hong and colleagues used multilevel PS stratification to estimate the effects of kindergarten retention evaluated balance within the 7-15 strata used in each study rather than within each school (Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007, 2008). Although the simulation considered the number of units within a cluster as the major factor for indicating which balance measures should be use, it is likely that the ratio of treated to untreated units is just as important. This could be another condition to examine in future research.

One limitation of using the ECLS-K dataset to define the parameters to be used in the generation of the simulation data was that the variances of the covariates were very similar between treatment and control groups prior to matching. Because the variance ratios of the covariates were fixed across the simulation conditions, this meant that variances of the covariates were already well balanced prior to matching. Even though the results suggested that the ASB performed better than the VR for both of the study's outcomes, variance ratios should not be completely ruled out as a balance measure for

multilevel PS matching. It is not clear if variance ratios would be the preferred method for datasets that have greater differences in the variances between treatment and control groups prior to matching.

Another challenge with using the ECLS-K parameters was that there were very small covariances among the cluster-level variables, making data generation difficult in some conditions (Table 1). The PS model estimates for all of the conditions converged; however, in some conditions, error messages indicated that "at least one element of the gradient is greater than 1e-3." This was most problematic for the condition with the lowest ICCs and the clusters with 10 units each; in this condition, the error message occurred in 56% of replications for the RIS propensity score model. According to Kiernan, Tao, and Gibbs (2012), this warning in SAS is common and typically is not a concern if the gradient values are reasonably small. These authors suggest that to confirm that estimates are not problematic, one can change the maximum gradient to 0 and compare the estimates and standard errors to the original results. Upon following these procedures, there was little to no change in the results.

The simulation also focused on four variables that were strongly correlated with kindergarten retention rather than the fuller set of covariates used in the PS model of the empirical analysis. Using a smaller number of covariates for a simulation than with real data is common in studies of multilevel PS methods (Thoemmes & West, 2011; Rickles & Seltzer, 2014; Arpino & Cannas, 2016). Thoemmes and West argued that increasing the number of covariates should not affect the performance of PS estimation methods given that models have been correctly specified. Increasing the number of covariates would have diminishing returns according to the proportion of variance that they explain.

The empirical analyses provided a means of examining the application of multilevel PS balance measures with real data using a larger set of covariates than in the simulation.

Another limitation is that the simulation conditions estimated treatment effects using just one method, calculating the difference in treatment and control means across the pooled sample. Although the pooled ASB was the best overall balance measure for predicting TE estimate bias based on that calculation, it is likely that a within-cluster balance measure would be better for estimating a separate treatment effect for each cluster. In an applied study using multilevel PS matching, Kim and Seltzer (2007) separately estimated balance and a TE estimate for each school. Because they estimated separate treatment effects for each cluster, it was more important to achieve balance within each cluster than to achieve balance across the pooled sample.

Additionally, although the calculation of a TE estimate based on the difference in means is common in practice, it is not a doubly robust method. In a doubly robust design, the use of PS methods paired with a TE model that controls for remaining baseline differences reduces TE estimation bias (Robins et al., 1994; Funk et al., 2011). As shown in the empirical illustration, the TE estimates from the different PS matched samples converge when the covariates are used in both the PS model and the TE model. It would be useful to know the extent to which the balance measure is important for selecting the correct model and estimating the treatment effect with the use of doubly robust methods that account for baseline differences in both models.

One might also argue that the premise of the first outcome, selecting the correctly specified model, is flawed, because it depends on the philosophy that the researcher should select only one PS model and estimate the TE according to that model. This is

generally the approach methodological researchers have taken when assessing different PS models, matching methods, and balance measures. However, Burnham and Anderson (2002, 2004) present a different philosophy of modeling that could be applied to PS methods. They argue that researchers do not have to choose just one model from a set of plausible models. Instead, researchers can compute an average parameter estimate across the tested models, with each model weighted according to the plausibility of being correct. These weights can be computed according to the AIC model fit statistic, or if bootstrapping is conducted, according to the estimated model selection frequencies. Burnham and Anderson (2004) show that this approach, which they call "multimodel inference," increases precision and reduces bias of the parameter estimate compared to selection of just one model. Multimodel inference could be extended to PS modeling to further reduce TE estimate bias. Instead of selecting the one PS model that leads to the most balanced samples, the researcher could estimate the treatment effects for all of the matched samples derived from the different PS models and matching methods tested and then calculate an average across those estimates. This reduces the risk that the researcher would select the PS model resulting in an outlying TE estimate compared to the TE estimates resulting from other potential PS models. Another way of addressing this concern would be through sensitivity tests, which some consider as a separate step in PS methods but was not evaluated in the simulation (Caliendo & Kopeinig, 2008).

The assumption that a researcher should select only one true PS model also had implications for the empirical illustrations. In the ECLS-K analysis, the TE estimates varied widely according to which PS model was selected due to the small number of matched pairs resulting from each PS model. King and Nielsen (forthcoming) argue that

model dependence, the problem of obtaining different TE estimates based on the model selected, is especially problematic when propensity scores are used for matching. As more observations are removed from the sample through matching, model dependence increases, and thus bias and imbalance also increase. They argue that because researchers have their own agendas and biases, they are likely to choose the model that will confirm their hypotheses, leading them to selecting a model resulting in an outlying TE estimate rather than the one that is most likely true. Although researchers in the PS literature advise that propensity score modeling should be completed as a first step before estimating treatment effects, they do not think that this is done in practice. They suspect that either consciously or unconsciously, researchers are selecting the methods that conform to their expected results. However, it seems that Burnham and Anderson's (2004) approach would help mitigate these concerns. If researchers average all possible TE estimates rather than picking the one that matches their hypotheses, it would likely result in less biased results and more accurate conclusions. Future research on PS methods should apply Burnham and Anderson's approach to examine its effects on TE estimate bias.

6.3 Summary

In interpreting the results of the study, one must be aware of the conditions that were not tested and the resulting limitations. Examining these limitations also sheds light on the gaps remaining in the literature on multilevel PS methods and directions for future research. For example, the simulation did not test all possible balance measures, PS methods, ICCs of the covariates, sample sizes, cluster sizes, or ratios between the number of treatment and control units. Moreover, the study did not consider other types of TE

estimation methods that could lead to less biased results, including estimation of separate treatment effects for each cluster (Kim & Seltzer, 2007), estimation of treatment effects that include covariate regression adjustment (Robins et al., 1994, Funk et al., 2011), and estimation of an average treatment effect across plausible PS and TE models (Burnham & Anderson, 2002, 2004). Assessing the correlation between the balance measures and these different types of TE estimates will be an important next step.

Nevertheless, the study increased the knowledge of PS methods and balance measures in multilevel settings. It demonstrated how using the best fitting PS model (in this case, the RIS model) often led to more biased TE estimates compared to using a simpler SL or RI model that results in more balanced samples. This provided justification for making recommendations to researchers based on the results from the second outcome, the correlation between balance measures and pooled TE estimate bias. These results suggest that in most cases the pooled ASB will have the highest correlation with the TE estimate, with the exception of large cluster sizes. With cluster sizes of 400, the matched samples had such low levels of imbalance across all balance measures that none of the balance measures had strong, positive relations with TE estimate bias. In this case, selecting the best PS model and matching method was less critical because all would lead to nearly the same TE estimate. The results also suggest that when averaging the balance results across many covariates, researchers should weigh them according to their likely influence on the outcome measure. Future studies that investigate balance measures for multilevel PS methods will determine how well these results hold with different data generating parameters and TE estimation methods.

Appendix

The appendix provides supplementary tables, including the covariance parameters from the simulation, descriptive statistics for the treatment effect estimate bias and balance measures and outcomes for all of the conditions tested in the simulation.

Table A1

Covariance structure of each of the three conditions with different average intracluster correlations for the unit-level covariates

	X_R school mean	X_M school mean	X_A school mean	*W*
Average ICC=.08				
X_R school mean	.048			
X_M school mean	.038	.046		
X_A school mean	.036	.041	1.276	
W	-.020	-.034	.008	1.288
Average ICC=.15				
X_R school mean	.097			
X_M school mean	.076	.091		
X_A school mean	.072	.082	2.553	
W	-.040	-.068	.016	2.575
Average ICC=.43				
X_R school mean	.386			
X_M school mean	.305	.365		
X_A school mean	.287	.328	10.211	
W	-.160	-.271	.065	10.301

Note. ICC=intracluster correlation.

The covariance structure for the condition with an average ICC of .27 is reported in Table 1.

Table A2.

Absolute treatment effect estimate bias by ICC, cluster size, propensity score model, and matching method.

Average ICC	Cluster size	Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No match
		P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	
.08	10	.157	.143	.117	.158	.145	.120	.096	.091	.065	.109	.105	.051	.121	.117	.052	.515
	25	.086	.081	.067	.090	.083	.066	.050	.051	.028	.110	.103	.022	.118	.110	.023	.540
	100	.030	.029	.024	.031	.029	.024	.018	.029	.013	.110	.099	.010	.118	.106	.010	.550
	400	.015	.014	.013	.016	.015	.013	.010	.022	.012	.111	.099	.009	.119	.105	.009	.554
.15	10	.170	.156	.118	.165	.159	.118	.100	.099	.063	.111	.106	.053	.122	.118	.050	.520
	25	.089	.080	.069	.087	.081	.067	.049	.051	.029	.110	.101	.022	.122	.112	.023	.542
	100	.031	.031	.027	.031	.031	.027	.020	.033	.017	.109	.097	.012	.121	.108	.012	.551
	400	.017	.017	.015	.018	.018	.015	.012	.026	.015	.109	.095	.011	.122	.106	.012	.554
.27	10	.169	.150	.120	.163	.148	.117	.098	.096	.064	.108	.101	.052	.130	.122	.053	.519
	25	.090	.084	.067	.091	.087	.071	.052	.058	.030	.108	.098	.024	.128	.119	.024	.542
	100	.035	.035	.029	.035	.034	.029	.023	.035	.019	.107	.093	.015	.128	.112	.015	.552
	400	.020	.021	.018	.021	.020	.018	.014	.030	.019	.107	.092	.015	.127	.109	.015	.554
.42	10	.172	.162	.126	.175	.164	.126	.105	.104	.066	.109	.099	.051	.144	.134	.052	.528
	25	.092	.088	.074	.093	.091	.076	.054	.059	.033	.102	.088	.025	.139	.126	.027	.544
	100	.040	.040	.034	.040	.040	.034	.025	.039	.024	.104	.087	.019	.138	.119	.020	.556
	400	.024	.025	.023	.025	.025	.023	.018	.033	.024	.103	.085	.020	.138	.117	.021	.558

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A3.

Pooled absolute standardized bias with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.244	.232	.189	.248	.231	.192	.156	.143	.119	.038	.038	.090	.070	.064	.091	.544
	25	.119	.117	.098	.118	.118	.098	.068	.063	.044	.025	.025	.040	.060	.051	.039	.554
	100	.035	.037	.030	.035	.038	.029	.024	.023	.015	.014	.019	.017	.052	.040	.017	.560
	400	.013	.010	.008	.013	.010	.008	.013	.009	.009	.010	.018	.011	.048	.037	.011	.561
.15	10	.247	.232	.186	.244	.230	.186	.162	.146	.115	.037	.039	.096	.086	.079	.094	.552
	25	.119	.116	.098	.119	.115	.096	.067	.062	.045	.027	.025	.039	.076	.065	.038	.565
	100	.035	.037	.030	.034	.038	.031	.023	.024	.016	.015	.018	.017	.067	.053	.017	.569
	400	.013	.011	.008	.013	.011	.008	.013	.009	.009	.010	.018	.011	.064	.051	.011	.571
.27	10	.244	.223	.181	.246	.221	.181	.153	.140	.114	.039	.039	.094	.105	.096	.093	.564
	25	.115	.114	.093	.119	.117	.094	.067	.065	.045	.025	.025	.040	.095	.082	.039	.576
	100	.036	.038	.029	.036	.038	.029	.023	.022	.016	.014	.018	.017	.089	.072	.017	.582
	400	.013	.011	.008	.013	.011	.008	.013	.009	.009	.010	.018	.011	.085	.069	.011	.583
.42	10	.247	.231	.184	.248	.232	.191	.158	.146	.114	.039	.039	.093	.136	.124	.092	.585
	25	.113	.113	.096	.113	.114	.098	.068	.061	.045	.026	.025	.040	.127	.112	.040	.595
	100	.036	.038	.030	.035	.038	.030	.023	.021	.016	.014	.017	.016	.118	.098	.016	.601
	400	.013	.011	.008	.013	.011	.008	.013	.009	.009	.010	.016	.011	.115	.094	.011	.601

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A4.

Pooled absolute standardized bias with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.309	.290	.237	.310	.293	.243	.187	.173	.139	.032	.033	.106	.040	.038	.109	.721
	25	.139	.140	.122	.139	.141	.121	.075	.072	.050	.022	.023	.045	.029	.028	.044	.741
	100	.036	.039	.033	.036	.039	.033	.022	.024	.018	.012	.019	.021	.021	.022	.021	.747
	400	.012	.010	.009	.012	.010	.009	.016	.009	.011	.010	.019	.016	.018	.020	.015	.751
.15	10	.315	.296	.234	.309	.296	.235	.196	.176	.133	.033	.035	.114	.043	.042	.110	.722
	25	.141	.140	.123	.141	.138	.121	.072	.071	.051	.023	.023	.044	.033	.031	.043	.741
	100	.035	.039	.034	.034	.039	.034	.021	.026	.019	.013	.019	.021	.025	.024	.021	.747
	400	.012	.011	.009	.013	.011	.009	.016	.009	.011	.010	.019	.016	.022	.023	.016	.749
.27	10	.310	.285	.237	.309	.281	.233	.182	.167	.132	.035	.035	.111	.048	.045	.109	.718
	25	.134	.134	.115	.138	.138	.118	.072	.075	.050	.022	.024	.046	.037	.035	.045	.737
	100	.036	.040	.033	.037	.039	.032	.021	.024	.017	.013	.018	.021	.030	.028	.020	.745
	400	.012	.011	.009	.012	.011	.009	.017	.009	.011	.010	.020	.015	.027	.027	.015	.747
.42	10	.309	.295	.237	.309	.294	.241	.186	.175	.133	.036	.036	.109	.055	.050	.106	.719
	25	.129	.132	.119	.129	.134	.123	.070	.068	.052	.023	.023	.047	.045	.043	.045	.733
	100	.035	.040	.033	.035	.040	.033	.022	.022	.018	.014	.018	.020	.037	.034	.019	.742
	400	.012	.011	.009	.013	.011	.009	.017	.009	.011	.010	.019	.015	.035	.033	.014	.744

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A5.

Percentage of covariates with pooled absolute standardized bias >.1 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.772	.767	.559	.784	.761	.571	.649	.614	.456	.047	.047	.369	.215	.200	.372	0.959
	25	.557	.539	.452	.553	.553	.449	.246	.212	.124	.006	.003	.108	.226	.181	.100	0.973
	100	.013	.021	.016	.015	.022	.014	.001	.001	.001	.000	.000	.001	.248	.187	.001	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.212	.000	1.000
.15	10	.780	.751	.574	.788	.757	.570	.664	.625	.440	.044	.042	.388	.250	.242	.372	0.990
	25	.557	.534	.449	.552	.530	.443	.225	.197	.141	.005	.002	.110	.249	.240	.093	0.999
	100	.017	.018	.016	.013	.017	.020	.000	.000	.001	.000	.000	.001	.250	.249	.002	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.250	.000	1.000
.27	10	.789	.760	.555	.798	.757	.548	.644	.615	.426	.050	.045	.376	.266	.261	.371	0.999
	25	.552	.537	.420	.552	.547	.429	.239	.222	.134	.005	.002	.105	.251	.250	.097	1.000
	100	.013	.021	.014	.014	.017	.018	.001	.001	.001	.000	.000	.001	.250	.250	.001	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.250	.000	1.000
.42	10	.812	.765	.573	.812	.762	.568	.662	.618	.434	.055	.054	.373	.270	.264	.349	1.000
	25	.529	.522	.437	.531	.525	.444	.238	.188	.147	.003	.002	.095	.250	.251	.107	1.000
	100	.014	.021	.018	.011	.025	.018	.002	.001	.001	.000	.000	.001	.250	.250	.002	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.250	.000	1.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A6.

Percentage of covariates with pooled absolute standardized bias >.1 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.919	.901	.706	.923	.908	.717	.795	.750	.557	.018	.024	.445	.058	.058	.459	0.990
	25	.683	.691	.568	.692	.704	.553	.303	.275	.119	.001	.001	.097	.055	.044	.092	0.993
	100	.013	.019	.008	.014	.019	.006	.000	.001	.000	.000	.000	.000	.060	.045	.000	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.051	.000	1.000
.15	10	.924	.907	.721	.919	.915	.732	.795	.769	.536	.018	.014	.472	.068	.066	.442	0.997
	25	.700	.694	.580	.700	.688	.561	.264	.254	.134	.001	.000	.108	.060	.058	.083	1.000
	100	.014	.015	.009	.012	.012	.012	.000	.000	.000	.000	.000	.000	.061	.060	.000	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.061	.000	1.000
.27	10	.913	.906	.725	.925	.899	.701	.778	.749	.507	.021	.025	.454	.069	.069	.447	1.000
	25	.673	.670	.528	.661	.686	.542	.271	.288	.122	.001	.000	.109	.061	.060	.098	1.000
	100	.010	.017	.008	.012	.013	.009	.001	.001	.000	.000	.000	.000	.061	.061	.000	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.061	.000	1.000
.42	10	.930	.903	.737	.926	.905	.721	.771	.740	.539	.033	.037	.449	.069	.065	.411	1.000
	25	.623	.654	.549	.639	.669	.574	.252	.238	.146	.001	.000	.097	.061	.061	.100	1.000
	100	.012	.016	.011	.012	.022	.011	.002	.001	.000	.000	.000	.000	.061	.061	.000	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.061	.000	1.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A7.

Percentage of covariates with pooled absolute standardized bias >.25 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.463	.427	.327	.463	.425	.322	.192	.146	.159	.000	.000	.079	.051	.035	.078	0.764
	25	.054	.052	.063	.055	.059	.077	.001	.000	.004	.000	.000	.002	.027	.006	.001	0.752
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	0.750
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	0.750
.15	10	.472	.437	.304	.466	.424	.305	.214	.152	.138	.000	.000	.093	.118	.086	.091	0.813
	25	.050	.054	.063	.051	.047	.057	.001	.000	.002	.000	.000	.002	.108	.036	.001	0.782
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.071	.002	.000	0.753
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.036	.000	.000	0.750
.27	10	.454	.409	.308	.458	.393	.301	.182	.143	.141	.000	.000	.096	.194	.169	.091	0.920
	25	.042	.045	.058	.049	.054	.061	.001	.002	.003	.000	.000	.002	.213	.152	.003	0.931
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.248	.110	.000	0.966
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.055	.000	0.997
.42	10	.458	.427	.305	.459	.438	.322	.199	.156	.135	.001	.000	.084	.247	.238	.096	0.989
	25	.044	.047	.064	.036	.045	.072	.001	.000	.003	.000	.000	.001	.250	.247	.003	1.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.250	.000	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.250	.250	.000	1.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A8.

Percentage of covariates with pooled absolute standardized bias >.25 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.674	.611	.421	.645	.618	.412	.262	.209	.166	.000	.000	.073	.012	.008	.078	0.943
	25	.066	.067	.071	.075	.080	.085	.000	.000	.003	.000	.000	.000	.006	.001	.000	0.940
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	0.939
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	0.939
.15	10	.676	.628	.396	.660	.620	.393	.307	.214	.145	.000	.000	.100	.029	.021	.089	0.955
	25	.066	.072	.073	.067	.057	.067	.001	.000	.000	.000	.000	.000	.026	.009	.000	0.947
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.017	.000	.000	0.940
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.009	.000	.000	0.939
.27	10	.649	.595	.405	.658	.585	.386	.248	.193	.141	.000	.000	.097	.047	.041	.087	0.981
	25	.055	.062	.063	.069	.070	.072	.001	.003	.001	.000	.000	.000	.052	.037	.001	0.983
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.060	.027	.000	0.992
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.013	.000	0.999
.42	10	.655	.627	.393	.649	.619	.411	.276	.229	.136	.000	.000	.082	.060	.058	.093	0.998
	25	.057	.058	.074	.043	.057	.085	.001	.000	.001	.000	.000	.000	.061	.060	.001	1.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.061	.000	1.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.061	.061	.000	1.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A9.

Pooled variance ratio with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

Average ICC	Cluster size	Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No match
		P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	
.08	10	.171	.168	.205	.175	.171	.211	.160	.149	.187	.111	.108	.157	.102	.102	.159	.090
	25	.110	.101	.106	.111	.100	.105	.118	.101	.094	.083	.076	.083	.073	.069	.082	.065
	100	.054	.047	.045	.054	.047	.045	.091	.063	.051	.065	.054	.044	.052	.046	.044	.044
	400	.025	.022	.021	.026	.022	.021	.081	.047	.038	.060	.047	.033	.043	.036	.033	.039
.15	10	.170	.166	.213	.171	.169	.215	.162	.151	.181	.113	.108	.156	.104	.103	.155	.090
	25	.109	.101	.106	.111	.102	.108	.119	.104	.095	.087	.079	.084	.075	.071	.084	.064
	100	.051	.046	.043	.052	.046	.042	.093	.064	.049	.073	.059	.044	.054	.046	.044	.046
	400	.026	.023	.021	.027	.023	.022	.085	.050	.038	.066	.053	.032	.046	.037	.032	.042
.27	10	.171	.165	.205	.176	.168	.207	.164	.155	.183	.120	.115	.156	.107	.103	.153	.093
	25	.108	.101	.107	.110	.101	.109	.121	.103	.095	.096	.087	.082	.075	.071	.084	.067
	100	.053	.046	.044	.054	.046	.043	.094	.065	.049	.081	.066	.043	.056	.047	.042	.050
	400	.026	.022	.020	.026	.022	.021	.086	.052	.038	.075	.060	.032	.049	.039	.032	.045
.42	10	.185	.172	.207	.182	.172	.206	.167	.158	.184	.124	.124	.156	.107	.104	.155	.098
	25	.107	.101	.109	.111	.101	.110	.123	.105	.097	.104	.092	.083	.079	.076	.083	.073
	100	.054	.047	.044	.053	.046	.044	.098	.066	.048	.089	.073	.043	.061	.051	.042	.056
	400	.025	.022	.021	.026	.022	.021	.089	.052	.037	.085	.068	.033	.054	.043	.032	.051

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A10.

Pooled variance ratio with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.187	.186	.256	.195	.188	.265	.180	.162	.231	.113	.110	.188	.097	.098	.191	.097
	25	.119	.109	.130	.120	.107	.128	.140	.112	.118	.091	.083	.100	.070	.066	.099	.070
	100	.059	.051	.054	.059	.051	.055	.118	.071	.064	.078	.061	.051	.046	.041	.051	.048
	400	.027	.024	.025	.028	.024	.024	.115	.052	.045	.077	.056	.035	.038	.030	.035	.044
.15	10	.181	.182	.260	.188	.184	.270	.181	.170	.231	.122	.118	.197	.100	.100	.195	.096
	25	.119	.111	.135	.122	.113	.137	.138	.114	.119	.097	.085	.101	.069	.066	.102	.069
	100	.056	.050	.051	.056	.049	.050	.122	.073	.061	.092	.071	.051	.046	.041	.052	.050
	400	.030	.025	.026	.030	.026	.026	.122	.057	.046	.089	.069	.036	.039	.031	.035	.048
.27	10	.186	.177	.250	.197	.181	.254	.183	.168	.223	.133	.123	.193	.099	.096	.189	.098
	25	.118	.108	.131	.121	.108	.136	.144	.114	.117	.114	.101	.099	.068	.067	.102	.070
	100	.058	.051	.053	.060	.050	.052	.127	.076	.064	.107	.084	.050	.046	.040	.050	.054
	400	.029	.024	.024	.028	.024	.025	.125	.063	.048	.107	.084	.037	.040	.032	.036	.051
.42	10	.207	.185	.257	.204	.188	.256	.193	.178	.230	.142	.145	.190	.101	.099	.192	.108
	25	.120	.112	.134	.123	.110	.134	.148	.119	.123	.129	.112	.100	.067	.067	.103	.076
	100	.059	.051	.054	.056	.052	.054	.133	.079	.064	.124	.099	.052	.047	.042	.052	.060
	400	.027	.024	.025	.027	.023	.024	.131	.066	.050	.124	.100	.039	.041	.034	.037	.058

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A11.

Percentage of covariates with pooled variance ratio <.5 or >2 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.017	.015	.092	.018	.011	.099	.010	.004	.059	.000	.000	.028	.000	.000	.030	.000
	25	.001	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.15	10	.016	.012	.105	.016	.012	.100	.009	.005	.056	.000	.000	.021	.000	.000	.023	.000
	25	.001	.000	.001	.000	.000	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.27	10	.016	.015	.090	.012	.015	.099	.011	.005	.059	.001	.000	.023	.000	.000	.027	.000
	25	.000	.000	.001	.000	.000	.001	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.42	10	.021	.018	.098	.018	.015	.095	.012	.008	.065	.000	.000	.026	.000	.000	.029	.000
	25	.000	.000	.001	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A12.

Percentage of covariates with pooled variance ratio <.5 or >2 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	.019	.016	.112	.018	.009	.123	.009	.003	.069	.000	.000	.023	.000	.000	.025	.000
	25	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.15	10	.019	.013	.129	.023	.015	.127	.010	.006	.078	.000	.000	.022	.000	.000	.020	.000
	25	.001	.000	.003	.000	.000	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.27	10	.019	.018	.107	.017	.015	.118	.012	.007	.068	.001	.000	.021	.000	.000	.024	.000
	25	.000	.000	.000	.000	.000	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
.42	10	.028	.021	.118	.022	.014	.119	.017	.008	.080	.000	.000	.032	.000	.000	.035	.000
	25	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	100	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	400	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A13.

Within-cluster absolute standardized bias (mean across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	1.078	0.986	1.077	1.063	1.004	1.100	1.389	1.132	1.133	1.163	1.067	1.177	1.178	1.054	1.183	1.319
	25	0.717	0.656	0.785	0.719	0.650	0.790	0.878	0.754	0.887	0.823	0.777	0.829	0.825	0.769	0.832	1.194
	100	0.318	0.287	0.325	0.318	0.286	0.325	0.572	0.487	0.505	0.649	0.631	0.492	0.647	0.628	0.491	1.165
	400	0.142	0.126	0.130	0.142	0.125	0.129	0.446	0.377	0.379	0.587	0.582	0.374	0.582	0.578	0.374	1.148
.15	10	1.088	0.956	1.009	1.125	0.976	1.048	1.358	1.118	1.218	1.171	1.043	1.198	1.153	1.041	1.221	1.312
	25	0.720	0.648	0.830	0.721	0.645	0.809	0.881	0.756	0.888	0.834	0.780	0.834	0.826	0.775	0.828	1.203
	100	0.318	0.286	0.324	0.320	0.288	0.324	0.565	0.484	0.510	0.653	0.635	0.491	0.652	0.630	0.491	1.165
	400	0.142	0.126	0.130	0.141	0.126	0.130	0.444	0.378	0.380	0.588	0.583	0.375	0.584	0.581	0.375	1.147
.27	10	1.115	0.983	0.978	1.102	0.991	1.010	1.383	1.134	1.160	1.154	1.041	1.221	1.152	1.033	1.222	1.308
	25	0.718	0.641	0.789	0.721	0.645	0.775	0.885	0.764	0.906	0.834	0.782	0.843	0.825	0.773	0.843	1.204
	100	0.317	0.285	0.325	0.318	0.286	0.328	0.563	0.484	0.506	0.654	0.636	0.492	0.654	0.635	0.492	1.168
	400	0.143	0.129	0.132	0.144	0.129	0.133	0.443	0.378	0.381	0.589	0.584	0.376	0.588	0.586	0.375	1.146
.42	10	1.095	0.957	0.993	1.093	0.969	1.026	1.316	1.107	1.134	1.150	1.037	1.170	1.145	1.028	1.167	1.324
	25	0.718	0.654	0.779	0.737	0.650	0.794	0.883	0.767	0.899	0.828	0.784	0.849	0.825	0.776	0.845	1.215
	100	0.315	0.286	0.328	0.317	0.286	0.327	0.560	0.482	0.506	0.654	0.634	0.489	0.658	0.640	0.489	1.168
	400	0.145	0.131	0.135	0.145	0.132	0.136	0.442	0.380	0.385	0.592	0.583	0.379	0.595	0.596	0.378	1.150

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A14.

Within-cluster absolute standardized bias (mean across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	1.024	0.933	1.090	1.006	0.966	1.092	1.591	1.260	1.204	1.343	1.223	1.231	1.355	1.193	1.251	1.530
	25	0.675	0.611	0.735	0.676	0.604	0.742	1.022	0.856	0.962	0.990	0.930	0.907	0.992	0.918	0.912	1.418
	100	0.289	0.252	0.291	0.287	0.252	0.290	0.730	0.587	0.586	0.829	0.799	0.577	0.828	0.794	0.576	1.389
	400	0.126	0.108	0.112	0.126	0.108	0.112	0.620	0.487	0.474	0.782	0.762	0.472	0.778	0.756	0.472	1.371
.15	10	1.037	0.929	0.994	1.084	0.955	1.008	1.549	1.238	1.323	1.356	1.197	1.313	1.351	1.194	1.347	1.536
	25	0.677	0.606	0.772	0.675	0.600	0.754	1.025	0.856	0.954	1.008	0.932	0.905	0.999	0.927	0.895	1.430
	100	0.291	0.252	0.290	0.292	0.254	0.292	0.720	0.587	0.602	0.837	0.805	0.577	0.835	0.797	0.576	1.393
	400	0.126	0.109	0.113	0.126	0.108	0.112	0.616	0.488	0.475	0.784	0.762	0.474	0.779	0.758	0.474	1.368
.27	10	1.094	0.962	0.999	1.069	0.954	1.018	1.541	1.269	1.226	1.331	1.182	1.222	1.336	1.166	1.230	1.528
	25	0.673	0.605	0.764	0.680	0.607	0.748	1.033	0.869	0.980	1.008	0.936	0.917	1.000	0.923	0.917	1.433
	100	0.290	0.251	0.292	0.290	0.250	0.290	0.717	0.587	0.587	0.835	0.808	0.575	0.836	0.802	0.577	1.398
	400	0.128	0.111	0.115	0.129	0.111	0.114	0.615	0.488	0.477	0.784	0.764	0.474	0.783	0.763	0.473	1.369
.42	10	1.047	0.945	0.983	1.055	0.927	0.995	1.501	1.234	1.201	1.324	1.182	1.256	1.318	1.168	1.244	1.542
	25	0.671	0.603	0.741	0.687	0.606	0.755	1.022	0.857	0.969	1.000	0.936	0.921	0.995	0.923	0.917	1.445
	100	0.287	0.252	0.293	0.290	0.252	0.294	0.714	0.586	0.590	0.835	0.805	0.573	0.838	0.807	0.574	1.397
	400	0.129	0.112	0.117	0.129	0.114	0.118	0.613	0.488	0.479	0.788	0.764	0.476	0.792	0.774	0.475	1.374

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A15.

Within-cluster absolute standardized bias (median across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.773	0.660	0.858	0.754	0.675	0.894	0.931	0.786	0.824	0.868	0.766	0.766	0.878	0.764	0.770	1.186
	25	0.496	0.438	0.483	0.498	0.437	0.485	0.669	0.568	0.581	0.688	0.644	0.562	0.691	0.638	0.560	1.145
	100	0.230	0.199	0.215	0.227	0.200	0.215	0.466	0.405	0.407	0.572	0.560	0.400	0.569	0.556	0.400	1.157
	400	0.107	0.093	0.095	0.107	0.092	0.094	0.379	0.337	0.341	0.533	0.546	0.335	0.532	0.539	0.336	1.156
.15	10	0.745	0.640	0.799	0.782	0.670	0.869	0.939	0.770	0.852	0.864	0.765	0.755	0.853	0.759	0.756	1.171
	25	0.505	0.443	0.487	0.503	0.440	0.488	0.666	0.570	0.586	0.693	0.645	0.564	0.694	0.644	0.564	1.150
	100	0.230	0.199	0.215	0.230	0.200	0.216	0.465	0.403	0.406	0.576	0.562	0.400	0.577	0.559	0.401	1.156
	400	0.106	0.092	0.094	0.106	0.092	0.094	0.379	0.336	0.340	0.534	0.548	0.337	0.535	0.543	0.337	1.158
.27	10	0.773	0.664	0.778	0.762	0.661	0.835	0.907	0.771	0.858	0.856	0.767	0.761	0.853	0.761	0.760	1.171
	25	0.498	0.432	0.483	0.502	0.437	0.482	0.672	0.572	0.591	0.693	0.645	0.568	0.696	0.642	0.567	1.156
	100	0.231	0.199	0.215	0.231	0.200	0.216	0.464	0.402	0.407	0.576	0.567	0.401	0.577	0.564	0.400	1.161
	400	0.107	0.093	0.095	0.107	0.094	0.096	0.378	0.333	0.339	0.535	0.548	0.335	0.538	0.548	0.334	1.157
.42	10	0.773	0.659	0.820	0.776	0.680	0.850	0.909	0.765	0.836	0.851	0.771	0.755	0.852	0.759	0.771	1.181
	25	0.496	0.439	0.482	0.501	0.439	0.487	0.663	0.567	0.588	0.687	0.645	0.571	0.693	0.645	0.567	1.161
	100	0.230	0.200	0.217	0.232	0.200	0.217	0.459	0.399	0.404	0.576	0.566	0.397	0.580	0.569	0.397	1.162
	400	0.108	0.094	0.096	0.108	0.093	0.096	0.376	0.332	0.338	0.540	0.551	0.336	0.547	0.557	0.336	1.163

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A16.

Within-cluster absolute standardized bias (median across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.722	0.630	0.851	0.719	0.642	0.883	1.077	0.886	0.874	1.020	0.883	0.805	1.026	0.881	0.813	1.401
	25	0.463	0.407	0.457	0.465	0.405	0.455	0.797	0.662	0.638	0.850	0.790	0.620	0.851	0.779	0.617	1.385
	100	0.205	0.172	0.191	0.203	0.173	0.191	0.614	0.509	0.493	0.761	0.734	0.489	0.757	0.726	0.487	1.408
	400	0.095	0.079	0.082	0.094	0.078	0.081	0.540	0.441	0.435	0.731	0.741	0.430	0.730	0.727	0.430	1.408
.15	10	0.703	0.613	0.804	0.748	0.636	0.823	1.086	0.877	0.903	1.017	0.888	0.801	1.013	0.876	0.802	1.402
	25	0.469	0.410	0.464	0.467	0.404	0.464	0.791	0.665	0.638	0.862	0.789	0.619	0.862	0.786	0.619	1.395
	100	0.207	0.173	0.191	0.207	0.174	0.192	0.613	0.508	0.494	0.769	0.739	0.488	0.768	0.730	0.488	1.410
	400	0.094	0.078	0.081	0.094	0.079	0.081	0.538	0.439	0.433	0.732	0.743	0.433	0.731	0.728	0.431	1.408
.27	10	0.751	0.639	0.804	0.738	0.637	0.837	1.044	0.882	0.916	0.999	0.889	0.808	1.007	0.873	0.805	1.390
	25	0.465	0.401	0.464	0.470	0.406	0.458	0.803	0.670	0.646	0.859	0.790	0.629	0.866	0.783	0.627	1.402
	100	0.208	0.172	0.191	0.206	0.174	0.191	0.610	0.508	0.492	0.766	0.745	0.489	0.765	0.733	0.487	1.416
	400	0.095	0.079	0.082	0.095	0.079	0.082	0.537	0.435	0.433	0.733	0.744	0.429	0.733	0.732	0.428	1.408
.42	10	0.727	0.639	0.804	0.738	0.655	0.831	1.044	0.865	0.885	0.994	0.889	0.795	0.998	0.872	0.814	1.401
	25	0.463	0.408	0.461	0.466	0.406	0.465	0.783	0.653	0.642	0.848	0.788	0.626	0.857	0.781	0.622	1.405
	100	0.206	0.174	0.193	0.208	0.174	0.192	0.602	0.501	0.488	0.763	0.743	0.482	0.761	0.733	0.482	1.415
	400	0.095	0.079	0.082	0.096	0.079	0.082	0.533	0.433	0.431	0.738	0.747	0.431	0.742	0.736	0.429	1.418

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A17.

Percentage of clusters with pooled absolute standardized bias >.1 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.915	0.905	0.892	0.915	0.908	0.899	0.935	0.920	0.911	0.927	0.925	0.913	0.932	0.922	0.915	0.959
	25	0.887	0.868	0.876	0.884	0.866	0.875	0.914	0.899	0.902	0.915	0.911	0.899	0.917	0.911	0.900	0.964
	100	0.754	0.715	0.733	0.752	0.716	0.734	0.866	0.856	0.866	0.885	0.893	0.861	0.888	0.893	0.860	0.984
	400	0.512	0.446	0.457	0.512	0.444	0.457	0.808	0.830	0.836	0.871	0.884	0.834	0.872	0.884	0.833	0.992
.15	10	0.913	0.905	0.892	0.915	0.902	0.894	0.934	0.921	0.918	0.934	0.923	0.916	0.929	0.923	0.917	0.957
	25	0.887	0.871	0.880	0.886	0.870	0.879	0.914	0.899	0.900	0.916	0.911	0.899	0.914	0.911	0.901	0.965
	100	0.753	0.714	0.734	0.754	0.717	0.737	0.864	0.856	0.863	0.886	0.894	0.863	0.887	0.894	0.862	0.983
	400	0.507	0.441	0.453	0.511	0.441	0.454	0.807	0.828	0.836	0.870	0.884	0.834	0.872	0.889	0.833	0.992
.27	10	0.917	0.909	0.896	0.916	0.902	0.900	0.936	0.922	0.906	0.933	0.927	0.913	0.929	0.923	0.912	0.957
	25	0.885	0.867	0.878	0.885	0.867	0.875	0.914	0.900	0.901	0.915	0.911	0.902	0.914	0.909	0.901	0.965
	100	0.757	0.713	0.735	0.756	0.716	0.736	0.865	0.856	0.866	0.888	0.892	0.861	0.890	0.896	0.861	0.984
	400	0.512	0.446	0.456	0.513	0.448	0.460	0.807	0.826	0.837	0.869	0.881	0.835	0.876	0.889	0.832	0.992
.42	10	0.916	0.906	0.905	0.920	0.910	0.896	0.934	0.921	0.911	0.933	0.926	0.911	0.932	0.922	0.915	0.959
	25	0.885	0.868	0.879	0.884	0.870	0.880	0.914	0.899	0.902	0.916	0.913	0.902	0.915	0.911	0.899	0.966
	100	0.756	0.717	0.738	0.755	0.714	0.736	0.865	0.854	0.865	0.888	0.890	0.860	0.892	0.896	0.860	0.985
	400	0.513	0.449	0.461	0.516	0.444	0.460	0.809	0.824	0.835	0.871	0.882	0.833	0.882	0.893	0.831	0.991

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A18.

Percentage of clusters with pooled absolute standardized bias >.1 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.929	0.921	0.901	0.928	0.920	0.911	0.964	0.950	0.935	0.960	0.954	0.935	0.962	0.953	0.937	0.992
	25	0.896	0.876	0.888	0.894	0.872	0.885	0.949	0.931	0.931	0.955	0.949	0.928	0.956	0.950	0.928	0.997
	100	0.743	0.692	0.718	0.741	0.695	0.721	0.923	0.905	0.910	0.943	0.947	0.904	0.946	0.949	0.904	1.001
	400	0.471	0.389	0.404	0.471	0.388	0.405	0.904	0.894	0.897	0.943	0.949	0.895	0.946	0.952	0.895	1.008
.15	10	0.932	0.923	0.913	0.928	0.918	0.915	0.967	0.951	0.942	0.966	0.954	0.940	0.963	0.956	0.941	0.990
	25	0.895	0.879	0.891	0.895	0.877	0.889	0.947	0.934	0.929	0.956	0.951	0.926	0.954	0.950	0.928	0.997
	100	0.741	0.692	0.720	0.742	0.697	0.725	0.920	0.906	0.908	0.946	0.949	0.907	0.946	0.951	0.907	1.002
	400	0.468	0.386	0.402	0.471	0.386	0.403	0.900	0.891	0.896	0.942	0.949	0.894	0.946	0.958	0.894	1.008
.27	10	0.930	0.923	0.922	0.932	0.918	0.920	0.964	0.951	0.933	0.964	0.955	0.936	0.959	0.954	0.936	0.992
	25	0.893	0.874	0.891	0.893	0.874	0.887	0.948	0.932	0.929	0.956	0.950	0.931	0.955	0.949	0.931	0.997
	100	0.747	0.691	0.721	0.744	0.696	0.721	0.922	0.905	0.909	0.945	0.948	0.907	0.949	0.952	0.906	1.003
	400	0.473	0.390	0.403	0.475	0.391	0.405	0.900	0.890	0.897	0.941	0.947	0.895	0.950	0.959	0.894	1.008
.42	10	0.926	0.920	0.916	0.933	0.922	0.906	0.962	0.949	0.935	0.963	0.954	0.935	0.963	0.953	0.935	0.992
	25	0.895	0.876	0.893	0.892	0.877	0.892	0.946	0.932	0.930	0.955	0.952	0.929	0.954	0.952	0.927	0.998
	100	0.745	0.695	0.726	0.742	0.695	0.724	0.920	0.903	0.908	0.944	0.948	0.906	0.948	0.954	0.904	1.004
	400	0.473	0.391	0.406	0.475	0.387	0.408	0.903	0.888	0.895	0.942	0.946	0.894	0.954	0.964	0.893	1.007

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A19.

Percentage of clusters with pooled absolute standardized bias >.25 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.796	0.770	0.763	0.794	0.776	0.760	0.836	0.809	0.778	0.825	0.813	0.789	0.833	0.812	0.789	0.896
	25	0.719	0.683	0.702	0.721	0.681	0.700	0.788	0.752	0.758	0.792	0.781	0.751	0.792	0.780	0.750	0.909
	100	0.454	0.392	0.426	0.450	0.391	0.426	0.678	0.653	0.670	0.725	0.736	0.661	0.728	0.737	0.662	0.951
	400	0.154	0.121	0.128	0.153	0.120	0.125	0.563	0.582	0.599	0.680	0.709	0.590	0.681	0.713	0.591	0.972
.15	10	0.792	0.767	0.754	0.795	0.765	0.756	0.837	0.805	0.801	0.834	0.812	0.796	0.827	0.814	0.800	0.893
	25	0.725	0.687	0.710	0.721	0.685	0.707	0.788	0.752	0.759	0.793	0.781	0.753	0.789	0.780	0.753	0.910
	100	0.454	0.391	0.426	0.457	0.396	0.427	0.675	0.650	0.666	0.725	0.736	0.662	0.728	0.736	0.662	0.951
	400	0.151	0.118	0.125	0.151	0.117	0.125	0.564	0.582	0.600	0.679	0.712	0.593	0.684	0.719	0.593	0.971
.27	10	0.799	0.772	0.750	0.787	0.768	0.757	0.836	0.810	0.786	0.831	0.815	0.791	0.828	0.813	0.792	0.894
	25	0.720	0.680	0.708	0.724	0.681	0.705	0.787	0.752	0.762	0.793	0.783	0.756	0.791	0.780	0.756	0.910
	100	0.459	0.391	0.424	0.455	0.393	0.426	0.676	0.651	0.669	0.726	0.737	0.661	0.731	0.740	0.659	0.953
	400	0.152	0.119	0.126	0.154	0.120	0.126	0.563	0.575	0.597	0.680	0.712	0.590	0.690	0.724	0.588	0.971
.42	10	0.797	0.769	0.756	0.799	0.779	0.759	0.839	0.809	0.778	0.830	0.814	0.787	0.831	0.810	0.789	0.896
	25	0.722	0.683	0.705	0.720	0.683	0.705	0.786	0.753	0.760	0.793	0.782	0.752	0.792	0.782	0.752	0.912
	100	0.456	0.396	0.430	0.459	0.396	0.427	0.676	0.649	0.666	0.730	0.734	0.657	0.735	0.743	0.658	0.953
	400	0.154	0.119	0.126	0.156	0.118	0.126	0.563	0.574	0.595	0.683	0.710	0.590	0.699	0.732	0.589	0.971

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A20.

Percentage of clusters with pooled absolute standardized bias >.25 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.795	0.774	0.775	0.795	0.774	0.775	0.874	0.847	0.805	0.870	0.855	0.814	0.875	0.854	0.813	0.945
	25	0.714	0.671	0.701	0.712	0.669	0.699	0.841	0.799	0.793	0.854	0.843	0.785	0.854	0.841	0.787	0.958
	100	0.412	0.337	0.382	0.408	0.338	0.381	0.777	0.736	0.743	0.828	0.837	0.734	0.832	0.837	0.735	0.970
	400	0.115	0.080	0.085	0.113	0.079	0.083	0.725	0.700	0.705	0.822	0.832	0.699	0.825	0.839	0.699	0.973
.15	10	0.797	0.768	0.765	0.798	0.767	0.765	0.881	0.844	0.827	0.878	0.855	0.817	0.875	0.857	0.822	0.944
	25	0.715	0.674	0.709	0.712	0.670	0.705	0.837	0.801	0.793	0.859	0.844	0.788	0.853	0.842	0.789	0.959
	100	0.415	0.339	0.383	0.417	0.343	0.386	0.774	0.735	0.740	0.832	0.838	0.736	0.835	0.840	0.736	0.971
	400	0.113	0.077	0.084	0.114	0.075	0.084	0.724	0.698	0.707	0.823	0.834	0.703	0.828	0.845	0.702	0.973
.27	10	0.801	0.775	0.767	0.796	0.770	0.769	0.874	0.849	0.815	0.874	0.854	0.814	0.873	0.853	0.814	0.944
	25	0.710	0.668	0.708	0.716	0.670	0.702	0.839	0.799	0.797	0.858	0.846	0.794	0.857	0.842	0.795	0.959
	100	0.418	0.338	0.380	0.413	0.339	0.383	0.775	0.736	0.743	0.830	0.838	0.738	0.836	0.843	0.736	0.974
	400	0.113	0.077	0.085	0.115	0.078	0.084	0.724	0.693	0.703	0.822	0.834	0.699	0.833	0.854	0.697	0.974
.42	10	0.801	0.778	0.763	0.801	0.783	0.769	0.875	0.845	0.802	0.872	0.855	0.812	0.875	0.850	0.814	0.946
	25	0.714	0.671	0.708	0.709	0.671	0.704	0.834	0.800	0.793	0.856	0.845	0.790	0.858	0.847	0.788	0.962
	100	0.416	0.343	0.387	0.420	0.345	0.384	0.774	0.732	0.738	0.833	0.836	0.731	0.837	0.845	0.731	0.974
	400	0.115	0.077	0.085	0.118	0.076	0.085	0.725	0.689	0.701	0.825	0.834	0.699	0.841	0.864	0.698	0.974

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A21.

Within-cluster variance ratio (mean across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.697	0.679	0.693	0.700	0.684	0.693	0.693	0.666	0.670	0.656	0.642	0.682	0.660	0.642	0.681	0.572
	25	0.594	0.574	0.612	0.592	0.572	0.615	0.566	0.537	0.581	0.525	0.498	0.559	0.528	0.498	0.558	0.439
	100	0.387	0.364	0.385	0.385	0.365	0.386	0.363	0.336	0.351	0.351	0.311	0.340	0.355	0.312	0.339	0.292
	400	0.223	0.205	0.208	0.223	0.204	0.208	0.220	0.191	0.193	0.224	0.183	0.186	0.234	0.185	0.186	0.196
.15	10	0.699	0.683	0.698	0.697	0.679	0.701	0.691	0.665	0.674	0.658	0.640	0.679	0.655	0.639	0.678	0.572
	25	0.591	0.569	0.615	0.593	0.571	0.614	0.566	0.536	0.582	0.526	0.500	0.559	0.529	0.499	0.560	0.440
	100	0.385	0.365	0.385	0.386	0.366	0.386	0.362	0.336	0.352	0.349	0.309	0.338	0.357	0.312	0.338	0.291
	400	0.223	0.206	0.210	0.223	0.206	0.210	0.219	0.191	0.193	0.222	0.184	0.186	0.233	0.186	0.186	0.197
.27	10	0.696	0.677	0.692	0.695	0.675	0.699	0.688	0.664	0.675	0.654	0.640	0.680	0.655	0.638	0.679	0.573
	25	0.596	0.573	0.618	0.597	0.572	0.620	0.568	0.539	0.583	0.524	0.502	0.561	0.528	0.501	0.560	0.445
	100	0.386	0.366	0.386	0.387	0.366	0.386	0.362	0.337	0.353	0.351	0.309	0.340	0.359	0.314	0.339	0.290
	400	0.224	0.205	0.209	0.224	0.205	0.209	0.219	0.190	0.193	0.219	0.184	0.186	0.232	0.188	0.186	0.197
.42	10	0.697	0.675	0.689	0.694	0.674	0.687	0.688	0.659	0.670	0.654	0.637	0.679	0.649	0.637	0.677	0.572
	25	0.595	0.574	0.615	0.596	0.573	0.615	0.570	0.541	0.588	0.530	0.503	0.564	0.529	0.503	0.562	0.446
	100	0.386	0.366	0.388	0.387	0.366	0.387	0.360	0.337	0.353	0.352	0.309	0.339	0.360	0.314	0.338	0.289
	400	0.224	0.207	0.210	0.224	0.207	0.211	0.219	0.193	0.195	0.216	0.186	0.188	0.233	0.192	0.187	0.198

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A22.

Within-cluster variance ratio (mean across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.712	0.695	0.709	0.718	0.697	0.706	0.708	0.676	0.681	0.670	0.654	0.693	0.676	0.654	0.693	0.587
	25	0.607	0.582	0.620	0.604	0.582	0.624	0.580	0.545	0.587	0.538	0.509	0.564	0.542	0.508	0.563	0.453
	100	0.398	0.367	0.387	0.397	0.368	0.388	0.376	0.341	0.356	0.364	0.320	0.343	0.369	0.322	0.342	0.310
	400	0.232	0.205	0.208	0.231	0.204	0.208	0.232	0.194	0.195	0.238	0.193	0.187	0.250	0.196	0.187	0.223
.15	10	0.712	0.698	0.709	0.715	0.693	0.717	0.707	0.676	0.678	0.673	0.650	0.689	0.670	0.652	0.686	0.587
	25	0.605	0.577	0.623	0.607	0.580	0.621	0.578	0.544	0.589	0.541	0.510	0.567	0.543	0.510	0.567	0.455
	100	0.395	0.367	0.386	0.396	0.367	0.386	0.374	0.340	0.355	0.360	0.318	0.340	0.371	0.321	0.341	0.310
	400	0.232	0.207	0.210	0.232	0.207	0.210	0.231	0.195	0.196	0.235	0.194	0.187	0.249	0.198	0.187	0.223
.27	10	0.709	0.692	0.712	0.709	0.693	0.718	0.706	0.676	0.682	0.668	0.654	0.687	0.672	0.652	0.687	0.588
	25	0.608	0.582	0.626	0.612	0.584	0.630	0.581	0.547	0.591	0.538	0.514	0.567	0.543	0.513	0.567	0.461
	100	0.397	0.368	0.386	0.398	0.369	0.387	0.374	0.342	0.357	0.363	0.318	0.344	0.372	0.325	0.343	0.310
	400	0.233	0.206	0.208	0.233	0.206	0.209	0.232	0.194	0.196	0.231	0.195	0.188	0.249	0.201	0.188	0.224
.42	10	0.709	0.690	0.700	0.711	0.690	0.699	0.707	0.674	0.672	0.670	0.651	0.687	0.666	0.652	0.686	0.590
	25	0.608	0.583	0.625	0.610	0.580	0.622	0.585	0.551	0.595	0.544	0.515	0.569	0.544	0.516	0.568	0.463
	100	0.397	0.367	0.387	0.397	0.368	0.388	0.372	0.342	0.357	0.363	0.317	0.342	0.375	0.325	0.341	0.309
	400	0.233	0.207	0.210	0.233	0.208	0.210	0.232	0.197	0.198	0.228	0.196	0.189	0.251	0.205	0.189	0.224

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A23.

Within-cluster variance ratio (median across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.752	0.731	0.719	0.758	0.742	0.717	0.762	0.725	0.704	0.711	0.694	0.738	0.718	0.693	0.733	0.600
	25	0.629	0.601	0.656	0.627	0.598	0.660	0.594	0.551	0.605	0.536	0.501	0.578	0.542	0.499	0.576	0.425
	100	0.363	0.334	0.354	0.362	0.336	0.355	0.338	0.302	0.316	0.319	0.277	0.303	0.325	0.278	0.303	0.254
	400	0.194	0.175	0.177	0.193	0.175	0.177	0.191	0.160	0.161	0.188	0.156	0.156	0.198	0.156	0.155	0.161
.15	10	0.754	0.738	0.723	0.751	0.736	0.724	0.758	0.724	0.713	0.715	0.692	0.735	0.711	0.689	0.736	0.601
	25	0.627	0.594	0.657	0.630	0.596	0.655	0.593	0.548	0.610	0.537	0.502	0.577	0.541	0.501	0.580	0.425
	100	0.362	0.335	0.355	0.361	0.335	0.356	0.338	0.302	0.317	0.316	0.276	0.301	0.326	0.278	0.302	0.254
	400	0.193	0.175	0.178	0.193	0.175	0.178	0.190	0.160	0.160	0.186	0.158	0.155	0.197	0.158	0.155	0.162
.27	10	0.752	0.736	0.718	0.754	0.732	0.724	0.755	0.724	0.708	0.709	0.694	0.735	0.706	0.689	0.735	0.601
	25	0.631	0.599	0.662	0.634	0.598	0.663	0.594	0.553	0.612	0.533	0.505	0.580	0.539	0.501	0.580	0.430
	100	0.362	0.337	0.354	0.365	0.336	0.356	0.338	0.302	0.317	0.319	0.277	0.305	0.328	0.280	0.302	0.254
	400	0.194	0.174	0.176	0.194	0.174	0.176	0.190	0.159	0.160	0.183	0.158	0.155	0.197	0.158	0.155	0.161
.42	10	0.748	0.725	0.710	0.747	0.723	0.708	0.752	0.715	0.706	0.709	0.686	0.731	0.703	0.687	0.728	0.600
	25	0.630	0.600	0.657	0.632	0.599	0.657	0.596	0.554	0.617	0.541	0.505	0.583	0.538	0.504	0.580	0.433
	100	0.364	0.337	0.358	0.364	0.338	0.357	0.337	0.305	0.320	0.320	0.277	0.305	0.330	0.280	0.305	0.253
	400	0.194	0.176	0.178	0.194	0.176	0.179	0.191	0.162	0.163	0.181	0.159	0.156	0.199	0.159	0.156	0.161

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A24.

Within-cluster variance ratio (median across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.769	0.747	0.735	0.777	0.756	0.731	0.778	0.731	0.714	0.728	0.706	0.752	0.736	0.707	0.746	0.616
	25	0.641	0.609	0.662	0.639	0.608	0.668	0.609	0.557	0.608	0.550	0.511	0.579	0.556	0.509	0.578	0.438
	100	0.374	0.335	0.355	0.374	0.335	0.353	0.352	0.306	0.317	0.333	0.285	0.305	0.339	0.286	0.305	0.271
	400	0.202	0.174	0.177	0.200	0.174	0.176	0.203	0.162	0.162	0.201	0.166	0.156	0.213	0.167	0.155	0.185
.15	10	0.768	0.753	0.734	0.769	0.752	0.739	0.776	0.739	0.719	0.733	0.702	0.745	0.730	0.703	0.744	0.618
	25	0.642	0.601	0.662	0.646	0.606	0.659	0.604	0.554	0.614	0.553	0.512	0.582	0.556	0.511	0.585	0.441
	100	0.372	0.335	0.353	0.372	0.334	0.354	0.349	0.304	0.318	0.327	0.283	0.301	0.340	0.285	0.302	0.272
	400	0.202	0.175	0.178	0.201	0.174	0.177	0.202	0.163	0.162	0.197	0.168	0.156	0.213	0.168	0.155	0.184
.27	10	0.767	0.755	0.739	0.771	0.755	0.743	0.776	0.736	0.714	0.724	0.710	0.742	0.725	0.704	0.744	0.618
	25	0.645	0.609	0.669	0.650	0.610	0.674	0.607	0.560	0.617	0.548	0.516	0.583	0.557	0.514	0.584	0.447
	100	0.373	0.338	0.351	0.375	0.338	0.355	0.350	0.306	0.319	0.330	0.285	0.307	0.341	0.289	0.304	0.272
	400	0.202	0.174	0.175	0.202	0.174	0.175	0.202	0.163	0.162	0.195	0.169	0.156	0.213	0.168	0.155	0.183
.42	10	0.761	0.740	0.720	0.766	0.736	0.719	0.774	0.732	0.707	0.730	0.702	0.739	0.722	0.704	0.737	0.621
	25	0.644	0.607	0.664	0.645	0.605	0.661	0.613	0.562	0.621	0.556	0.516	0.584	0.554	0.517	0.584	0.450
	100	0.375	0.336	0.354	0.374	0.337	0.355	0.347	0.308	0.321	0.330	0.285	0.305	0.345	0.290	0.305	0.272
	400	0.202	0.175	0.177	0.202	0.176	0.178	0.203	0.165	0.165	0.192	0.169	0.157	0.216	0.170	0.156	0.182

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A25.

Percentage of clusters with pooled variance ratio <.5 or >2 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.759	0.733	0.747	0.759	0.741	0.749	0.750	0.719	0.717	0.711	0.693	0.733	0.714	0.692	0.730	0.609
	25	0.634	0.606	0.648	0.634	0.603	0.653	0.600	0.555	0.608	0.542	0.503	0.581	0.547	0.499	0.579	0.406
	100	0.316	0.278	0.310	0.316	0.280	0.313	0.275	0.231	0.256	0.250	0.185	0.237	0.259	0.188	0.235	0.160
	400	0.067	0.052	0.056	0.066	0.052	0.057	0.065	0.043	0.046	0.078	0.032	0.039	0.085	0.036	0.040	0.060
.15	10	0.757	0.735	0.742	0.758	0.730	0.757	0.747	0.715	0.717	0.714	0.691	0.726	0.707	0.689	0.722	0.609
	25	0.630	0.599	0.654	0.632	0.601	0.652	0.600	0.554	0.609	0.542	0.503	0.579	0.547	0.502	0.580	0.405
	100	0.313	0.277	0.310	0.313	0.279	0.310	0.273	0.230	0.258	0.249	0.183	0.234	0.260	0.189	0.235	0.161
	400	0.066	0.052	0.057	0.066	0.052	0.057	0.063	0.043	0.046	0.075	0.031	0.040	0.083	0.038	0.040	0.062
.27	10	0.751	0.727	0.740	0.750	0.725	0.752	0.744	0.716	0.722	0.708	0.691	0.731	0.712	0.689	0.729	0.610
	25	0.635	0.606	0.662	0.635	0.601	0.661	0.602	0.557	0.612	0.538	0.506	0.583	0.544	0.502	0.581	0.413
	100	0.317	0.278	0.310	0.319	0.278	0.312	0.273	0.232	0.258	0.251	0.181	0.239	0.266	0.189	0.237	0.159
	400	0.067	0.052	0.056	0.067	0.052	0.058	0.061	0.042	0.046	0.072	0.032	0.040	0.082	0.042	0.040	0.064
.42	10	0.757	0.731	0.741	0.753	0.726	0.730	0.744	0.709	0.715	0.708	0.687	0.731	0.701	0.685	0.728	0.609
	25	0.636	0.606	0.654	0.638	0.603	0.655	0.602	0.560	0.615	0.548	0.507	0.584	0.544	0.504	0.582	0.415
	100	0.317	0.283	0.316	0.318	0.282	0.316	0.272	0.233	0.261	0.253	0.184	0.236	0.269	0.193	0.236	0.160
	400	0.067	0.052	0.057	0.066	0.052	0.056	0.061	0.042	0.046	0.069	0.031	0.041	0.085	0.045	0.041	0.065

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A26.

Percentage of clusters with pooled variance ratio <.5 or >2 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.774	0.748	0.761	0.780	0.752	0.762	0.766	0.726	0.729	0.725	0.705	0.746	0.730	0.703	0.743	0.625
	25	0.648	0.615	0.655	0.647	0.613	0.662	0.615	0.562	0.612	0.556	0.513	0.582	0.562	0.509	0.582	0.422
	100	0.327	0.276	0.306	0.329	0.276	0.308	0.290	0.233	0.257	0.266	0.191	0.236	0.276	0.196	0.235	0.182
	400	0.073	0.049	0.053	0.071	0.049	0.053	0.073	0.044	0.047	0.087	0.036	0.039	0.097	0.041	0.039	0.078
.15	10	0.770	0.751	0.751	0.778	0.743	0.770	0.764	0.727	0.719	0.730	0.702	0.734	0.722	0.703	0.730	0.623
	25	0.645	0.607	0.662	0.648	0.611	0.658	0.611	0.559	0.614	0.559	0.512	0.584	0.564	0.513	0.586	0.423
	100	0.322	0.273	0.303	0.324	0.274	0.302	0.287	0.231	0.257	0.263	0.191	0.231	0.278	0.199	0.233	0.185
	400	0.071	0.050	0.055	0.073	0.051	0.055	0.070	0.043	0.046	0.084	0.035	0.039	0.096	0.044	0.040	0.081
.27	10	0.765	0.742	0.766	0.764	0.744	0.771	0.764	0.728	0.733	0.723	0.704	0.737	0.731	0.705	0.735	0.626
	25	0.649	0.615	0.669	0.651	0.615	0.671	0.613	0.563	0.617	0.555	0.517	0.585	0.561	0.515	0.587	0.433
	100	0.329	0.274	0.304	0.330	0.276	0.306	0.288	0.234	0.257	0.264	0.190	0.239	0.282	0.201	0.235	0.187
	400	0.073	0.050	0.054	0.073	0.050	0.055	0.069	0.043	0.047	0.080	0.036	0.041	0.097	0.050	0.040	0.083
.42	10	0.771	0.749	0.754	0.774	0.745	0.746	0.768	0.728	0.716	0.726	0.705	0.738	0.719	0.701	0.738	0.630
	25	0.648	0.613	0.664	0.653	0.609	0.662	0.618	0.569	0.621	0.561	0.518	0.585	0.562	0.517	0.587	0.435
	100	0.327	0.275	0.308	0.327	0.277	0.309	0.283	0.234	0.261	0.265	0.192	0.235	0.289	0.204	0.234	0.185
	400	0.073	0.050	0.055	0.072	0.049	0.053	0.069	0.043	0.046	0.077	0.035	0.041	0.100	0.054	0.041	0.085

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A27.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Pooled absolute standardized bias with covariates equally weighted				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.00	0.00	0.08
	25	0.00	0.00	0.04
	100	0.06	0.07	0.10
	400	0.25	0.30	0.40
.15	10	0.00	0.00	0.06
	25	0.00	0.00	0.03
	100	0.07	0.07	0.10
	400	0.28	0.27	0.37
.27	10	0.00	0.00	0.07
	25	0.00	0.01	0.03
	100	0.08	0.07	0.08
	400	0.27	0.29	0.38
.42	10	0.00	0.00	0.08
	25	0.00	0.00	0.03
	100	0.08	0.06	0.10
	400	0.27	0.27	0.40

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A28.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Pooled absolute standardized bias with covariates weighted according to the strength of relation with the outcome				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.00	0.00	0.08
	25	0.00	0.00	0.04
	100	0.06	0.07	0.10
	400	0.25	0.30	0.40
.15	10	0.00	0.00	0.06
	25	0.00	0.00	0.03
	100	0.07	0.07	0.10
	400	0.28	0.27	0.37
.27	10	0.00	0.00	0.07
	25	0.00	0.01	0.03
	100	0.08	0.07	0.08
	400	0.27	0.29	0.38
.42	10	0.00	0.00	0.08
	25	0.00	0.00	0.03
	100	0.08	0.06	0.10
	400	0.27	0.27	0.40

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A29.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of covariates with pooled absolute standardized bias >.1 with covariates equally weighted				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.00	0.00	0.03
	25	0.00	0.00	0.01
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00
.15	10	0.00	0.00	0.04
	25	0.00	0.00	0.00
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00
.27	10	0.00	0.00	0.05
	25	0.00	0.00	0.01
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00
.42	10	0.00	0.00	0.05
	25	0.00	0.00	0.01
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A30.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of covariates with pooled absolute standardized bias >.1 with covariates weighted according to the strength of relation with the outcome				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.00	0.00	0.07
	25	0.00	0.00	0.02
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00
.15	10	0.00	0.00	0.06
	25	0.00	0.00	0.00
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00
.27	10	0.00	0.00	0.07
	25	0.00	0.00	0.02
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00
.42	10	0.00	0.00	0.07
	25	0.00	0.00	0.01
	100	0.00	0.00	0.00
	400	0.00	0.00	0.00

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A31.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of covariates with pooled absolute standardized bias >.25 with covariates equally weighted				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.000	0.000	0.010
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.15	10	0.000	0.000	0.018
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.27	10	0.000	0.000	0.008
	25	0.000	0.000	0.002
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.42	10	0.000	0.000	0.016
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A32.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of covariates with pooled absolute standardized bias >.25 with covariates weighted according to the strength of relation with the outcome				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.000	0.000	0.012
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.15	10	0.000	0.000	0.020
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.27	10	0.000	0.000	0.008
	25	0.000	0.000	0.002
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.42	10	0.000	0.000	0.022
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A33.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Pooled variance ratio with covariates equally weighted				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.080	0.086	0.152
	25	0.118	0.102	0.154
	100	0.298	0.248	0.228
	400	0.500	0.478	0.450
.15	10	0.094	0.094	0.142
	25	0.146	0.120	0.164
	100	0.342	0.302	0.246
	400	0.520	0.488	0.452
.27	10	0.088	0.076	0.144
	25	0.138	0.120	0.162
	100	0.378	0.298	0.226
	400	0.514	0.456	0.410
.42	10	0.068	0.066	0.142
	25	0.164	0.154	0.144
	100	0.356	0.314	0.240
	400	0.506	0.484	0.430

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A34.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Pooled variance ratio with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.088	0.084	0.154
	25	0.144	0.144	0.170
	100	0.244	0.224	0.254
	400	0.450	0.426	0.412
.15	10	0.116	0.106	0.166
	25	0.142	0.122	0.158
	100	0.276	0.258	0.252
	400	0.438	0.416	0.396
.27	10	0.110	0.110	0.168
	25	0.138	0.148	0.196
	100	0.284	0.220	0.250
	400	0.430	0.388	0.400
.42	10	0.076	0.102	0.170
	25	0.160	0.128	0.156
	100	0.272	0.248	0.232
	400	0.460	0.432	0.392

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A35.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of covariates with pooled variance ratio <.5 or >2 with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.000	0.000	0.006
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.15	10	0.000	0.000	0.002
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.27	10	0.000	0.000	0.006
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.42	10	0.000	0.000	0.008
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A36.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of covariates with pooled variance ratio <.5 or >2 with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.000	0.000	0.006
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.15	10	0.000	0.000	0.002
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.27	10	0.000	0.000	0.006
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000
.42	10	0.000	0.000	0.008
	25	0.000	0.000	0.000
	100	0.000	0.000	0.000
	400	0.000	0.000	0.000

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A37.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster absolute standardized bias (mean across clusters) with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.362	0.338	0.252
	25	0.486	0.466	0.340
	100	0.478	0.492	0.496
	400	0.486	0.460	0.462
.15	10	0.330	0.368	0.292
	25	0.472	0.444	0.288
	100	0.530	0.528	0.496
	400	0.492	0.494	0.492
.27	10	0.284	0.352	0.306
	25	0.496	0.514	0.370
	100	0.508	0.500	0.498
	400	0.538	0.516	0.496
.42	10	0.324	0.342	0.290
	25	0.512	0.444	0.366
	100	0.546	0.494	0.482
	400	0.488	0.494	0.508

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A38.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster absolute standardized bias (mean across clusters) with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.426	0.418	0.242
	25	0.516	0.494	0.428
	100	0.490	0.502	0.512
	400	0.514	0.478	0.466
.15	10	0.400	0.430	0.300
	25	0.514	0.476	0.384
	100	0.510	0.530	0.518
	400	0.486	0.504	0.476
.27	10	0.360	0.396	0.302
	25	0.536	0.534	0.408
	100	0.510	0.520	0.498
	400	0.482	0.510	0.488
.42	10	0.404	0.384	0.272
	25	0.516	0.486	0.440
	100	0.548	0.510	0.494
	400	0.502	0.490	0.522

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A39.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster absolute standardized bias (median across clusters) with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.344	0.364	0.216
	25	0.508	0.476	0.428
	100	0.474	0.494	0.506
	400	0.500	0.476	0.500
.15	10	0.360	0.430	0.264
	25	0.510	0.474	0.426
	100	0.504	0.506	0.480
	400	0.536	0.510	0.480
.27	10	0.316	0.350	0.248
	25	0.518	0.510	0.442
	100	0.482	0.528	0.528
	400	0.506	0.500	0.520
.42	10	0.340	0.376	0.252
	25	0.518	0.502	0.438
	100	0.492	0.512	0.490
	400	0.496	0.472	0.506

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A40.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster absolute standardized bias (median across clusters) with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.442	0.426	0.218
	25	0.512	0.496	0.442
	100	0.500	0.524	0.494
	400	0.502	0.466	0.472
.15	10	0.424	0.466	0.250
	25	0.478	0.460	0.482
	100	0.516	0.514	0.504
	400	0.526	0.526	0.506
.27	10	0.368	0.410	0.262
	25	0.526	0.530	0.446
	100	0.474	0.522	0.524
	400	0.520	0.524	0.512
.42	10	0.396	0.402	0.260
	25	0.486	0.498	0.492
	100	0.522	0.506	0.488
	400	0.518	0.502	0.498

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A41.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of clusters with pooled absolute standardized bias >.1 with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.258	0.284	0.198
	25	0.354	0.372	0.318
	100	0.486	0.516	0.496
	400	0.496	0.482	0.472
.15	10	0.284	0.250	0.202
	25	0.344	0.368	0.352
	100	0.498	0.496	0.518
	400	0.500	0.488	0.472
.27	10	0.262	0.254	0.188
	25	0.386	0.378	0.326
	100	0.472	0.532	0.494
	400	0.472	0.508	0.510
.42	10	0.286	0.276	0.160
	25	0.386	0.420	0.346
	100	0.494	0.454	0.454
	400	0.486	0.470	0.460

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A42.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of clusters with pooled absolute standardized bias >.1 with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.274	0.284	0.200
	25	0.454	0.426	0.350
	100	0.504	0.534	0.494
	400	0.496	0.498	0.544
.15	10	0.296	0.280	0.208
	25	0.444	0.442	0.380
	100	0.516	0.524	0.508
	400	0.502	0.500	0.504
.27	10	0.280	0.292	0.192
	25	0.460	0.464	0.360
	100	0.486	0.562	0.486
	400	0.514	0.518	0.494
.42	10	0.318	0.290	0.168
	25	0.448	0.462	0.368
	100	0.490	0.504	0.476
	400	0.512	0.500	0.502

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A43.

Proportion of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of clusters with pooled absolute standardized bias >.25 with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.284	0.320	0.188
	25	0.478	0.458	0.370
	100	0.468	0.504	0.476
	400	0.466	0.434	0.412
.15	10	0.308	0.292	0.216
	25	0.412	0.422	0.360
	100	0.496	0.512	0.482
	400	0.468	0.424	0.454
.27	10	0.262	0.278	0.218
	25	0.492	0.456	0.354
	100	0.464	0.514	0.482
	400	0.508	0.456	0.450
.42	10	0.308	0.342	0.204
	25	0.418	0.440	0.388
	100	0.512	0.488	0.450
	400	0.500	0.456	0.472

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A44.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of clusters with pooled absolute standardized bias >.25 with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.328	0.378	0.202
	25	0.480	0.498	0.418
	100	0.478	0.538	0.474
	400	0.486	0.486	0.476
.15	10	0.364	0.340	0.248
	25	0.468	0.462	0.406
	100	0.500	0.518	0.498
	400	0.528	0.462	0.504
.27	10	0.364	0.354	0.262
	25	0.516	0.498	0.398
	100	0.488	0.498	0.512
	400	0.510	0.518	0.482
.42	10	0.340	0.364	0.232
	25	0.468	0.498	0.428
	100	0.518	0.510	0.470
	400	0.544	0.466	0.500

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A45.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster variance ratio (mean across clusters) with covariates equally weighted				
Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.162	0.122	0.202
	25	0.012	0.014	0.060
	100	0.008	0.008	0.012
	400	0.210	0.012	0.026
.15	10	0.120	0.116	0.186
	25	0.016	0.012	0.058
	100	0.038	0.000	0.014
	400	0.166	0.018	0.018
.27	10	0.124	0.128	0.188
	25	0.006	0.008	0.042
	100	0.032	0.000	0.010
	400	0.164	0.032	0.024
.42	10	0.134	0.124	0.192
	25	0.016	0.010	0.050
	100	0.030	0.002	0.008
	400	0.156	0.022	0.016

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A46.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster variance ratio (mean across clusters) with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.172	0.164	0.190
	25	0.028	0.024	0.074
	100	0.022	0.018	0.040
	400	0.260	0.088	0.050
.15	10	0.138	0.140	0.186
	25	0.032	0.024	0.078
	100	0.060	0.022	0.028
	400	0.240	0.090	0.050
.27	10	0.144	0.148	0.182
	25	0.026	0.016	0.054
	100	0.044	0.014	0.026
	400	0.192	0.100	0.056
.42	10	0.172	0.132	0.184
	25	0.028	0.016	0.070
	100	0.050	0.028	0.026
	400	0.204	0.104	0.062

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A47.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster variance ratio (median across clusters) with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.176	0.168	0.208
	25	0.032	0.012	0.052
	100	0.026	0.018	0.024
	400	0.160	0.050	0.034
.15	10	0.156	0.144	0.226
	25	0.016	0.016	0.074
	100	0.044	0.012	0.028
	400	0.156	0.042	0.028
.27	10	0.150	0.150	0.208
	25	0.020	0.012	0.046
	100	0.058	0.008	0.030
	400	0.142	0.040	0.024
.42	10	0.170	0.156	0.216
	25	0.014	0.010	0.058
	100	0.036	0.016	0.020
	400	0.134	0.056	0.052

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A48.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Within-cluster variance ratio (median across clusters) with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.182	0.172	0.226
	25	0.064	0.026	0.064
	100	0.052	0.046	0.042
	400	0.200	0.142	0.070
.15	10	0.168	0.150	0.206
	25	0.046	0.046	0.112
	100	0.060	0.040	0.050
	400	0.210	0.126	0.066
.27	10	0.130	0.146	0.208
	25	0.044	0.028	0.086
	100	0.076	0.028	0.058
	400	0.182	0.126	0.076
.42	10	0.190	0.162	0.204
	25	0.030	0.030	0.078
	100	0.058	0.044	0.046
	400	0.194	0.146	0.098

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A49.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of clusters with pooled variance ratio <.5 or >2 with covariates equally weighted

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.134	0.154	0.154
	25	0.036	0.016	0.080
	100	0.018	0.006	0.022
	400	0.240	0.044	0.072
.15	10	0.130	0.136	0.180
	25	0.018	0.028	0.082
	100	0.028	0.006	0.024
	400	0.198	0.036	0.064
.27	10	0.144	0.154	0.172
	25	0.028	0.022	0.048
	100	0.024	0.000	0.020
	400	0.176	0.072	0.070
.42	10	0.126	0.142	0.130
	25	0.012	0.012	0.076
	100	0.036	0.002	0.020
	400	0.174	0.056	0.068

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A50.

Percentage of replications in which the RIS model was selected by ICC, cluster size, propensity score model, and matching method.

Balance measure: Percentage of clusters with pooled variance ratio <.5 or >2 with covariates weighted according to the strength of relation with the outcome

Average ICC	Cluster size	Pooled	Two-stage	Within cluster
.08	10	0.160	0.164	0.174
	25	0.062	0.030	0.108
	100	0.040	0.022	0.054
	400	0.252	0.108	0.110
.15	10	0.174	0.170	0.172
	25	0.050	0.050	0.108
	100	0.052	0.034	0.048
	400	0.246	0.094	0.108
.27	10	0.154	0.146	0.182
	25	0.032	0.040	0.062
	100	0.058	0.020	0.044
	400	0.230	0.124	0.126
.42	10	0.176	0.166	0.174
	25	0.036	0.028	0.092
	100	0.058	0.024	0.048
	400	0.222	0.114	0.114

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model.

Table A51.

Correlation between treatment effect estimate bias and pooled absolute standardized bias with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.49	0.49	0.52	0.52	0.50	0.53	0.51	0.48	0.45	-0.05	-0.07	0.48	0.11	0.09	0.47	0.62
	25	0.59	0.54	0.53	0.58	0.57	0.47	0.51	0.53	0.38	0.13	0.17	0.45	0.13	0.11	0.43	0.66
	100	0.54	0.53	0.40	0.51	0.54	0.38	0.27	0.47	0.13	0.11	0.08	0.07	0.13	0.19	0.10	0.64
	400	0.20	0.16	-0.04	0.20	0.10	-0.04	-0.35	0.14	-0.20	0.08	0.14	-0.28	0.24	0.16	-0.30	0.64
.15	10	0.53	0.55	0.49	0.54	0.52	0.45	0.61	0.53	0.40	0.04	0.04	0.51	0.11	0.18	0.51	0.69
	25	0.61	0.60	0.47	0.57	0.58	0.50	0.52	0.49	0.34	0.12	0.18	0.38	0.14	0.10	0.33	0.71
	100	0.48	0.59	0.42	0.51	0.58	0.47	0.14	0.50	0.12	0.14	0.13	0.06	0.22	0.12	-0.04	0.67
	400	0.15	0.19	-0.05	0.16	0.24	0.00	-0.41	0.32	-0.24	0.17	0.09	-0.32	0.20	0.16	-0.24	0.64
.27	10	0.56	0.58	0.62	0.55	0.51	0.58	0.49	0.40	0.42	0.08	0.02	0.46	0.08	0.07	0.50	0.65
	25	0.59	0.57	0.46	0.58	0.56	0.53	0.48	0.59	0.38	0.15	0.16	0.35	0.10	0.14	0.34	0.64
	100	0.48	0.48	0.40	0.45	0.48	0.39	0.18	0.39	0.07	0.15	0.24	0.05	0.20	0.17	0.09	0.66
	400	0.19	0.21	-0.04	0.16	0.10	-0.13	-0.40	0.17	-0.21	0.28	0.14	-0.26	0.21	0.16	-0.31	0.74
.42	10	0.52	0.51	0.50	0.57	0.56	0.53	0.52	0.56	0.47	0.06	-0.04	0.45	0.23	0.20	0.49	0.65
	25	0.54	0.59	0.51	0.58	0.56	0.54	0.44	0.51	0.32	0.13	0.02	0.19	0.22	0.20	0.26	0.70
	100	0.47	0.56	0.41	0.52	0.56	0.38	0.05	0.34	0.02	0.12	0.27	-0.07	0.27	0.26	0.05	0.70
	400	0.09	0.21	-0.06	0.23	0.19	-0.10	-0.53	0.01	-0.11	0.29	0.22	-0.25	0.23	0.23	-0.17	0.70

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A52.

Correlation between treatment effect estimate bias and pooled absolute standardized bias with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.69	0.69	0.72	0.70	0.70	0.76	0.66	0.65	0.50	0.01	0.01	0.58	0.12	0.15	0.56	0.81
	25	0.75	0.73	0.80	0.74	0.76	0.77	0.67	0.72	0.39	0.14	0.30	0.43	0.15	0.22	0.43	0.83
	100	0.70	0.71	0.69	0.65	0.75	0.71	0.32	0.61	-0.03	0.26	0.34	-0.21	0.21	0.43	-0.16	0.83
	400	0.23	0.26	-0.01	0.28	0.16	0.00	-0.52	0.24	-0.57	0.34	0.39	-0.68	0.41	0.42	-0.66	0.82
.15	10	0.70	0.72	0.74	0.69	0.68	0.74	0.72	0.68	0.51	0.08	0.14	0.58	0.06	0.22	0.57	0.84
	25	0.74	0.75	0.78	0.75	0.74	0.78	0.66	0.67	0.37	0.24	0.32	0.36	0.19	0.24	0.34	0.85
	100	0.67	0.75	0.71	0.66	0.74	0.73	0.17	0.63	-0.09	0.28	0.37	-0.28	0.36	0.37	-0.34	0.83
	400	0.19	0.19	-0.02	0.20	0.23	-0.01	-0.57	0.41	-0.58	0.32	0.40	-0.69	0.38	0.40	-0.65	0.82
.27	10	0.70	0.70	0.79	0.73	0.68	0.76	0.64	0.62	0.48	0.10	0.07	0.52	0.09	0.18	0.55	0.81
	25	0.76	0.77	0.75	0.77	0.77	0.77	0.67	0.72	0.35	0.15	0.26	0.22	0.16	0.26	0.21	0.84
	100	0.65	0.69	0.66	0.61	0.67	0.67	0.17	0.58	-0.13	0.25	0.42	-0.35	0.31	0.43	-0.33	0.83
	400	0.14	0.26	-0.03	0.20	0.20	-0.08	-0.61	0.33	-0.55	0.40	0.40	-0.69	0.36	0.41	-0.67	0.85
.42	10	0.69	0.69	0.74	0.72	0.73	0.74	0.68	0.72	0.54	0.07	0.05	0.48	0.17	0.16	0.51	0.80
	25	0.73	0.72	0.76	0.74	0.76	0.78	0.62	0.63	0.29	0.17	0.14	0.09	0.24	0.24	0.16	0.84
	100	0.62	0.73	0.66	0.67	0.73	0.64	-0.01	0.53	-0.22	0.23	0.45	-0.45	0.35	0.43	-0.34	0.84
	400	0.12	0.32	-0.07	0.23	0.32	-0.04	-0.69	0.15	-0.53	0.41	0.42	-0.71	0.39	0.42	-0.68	0.85

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A53.

Correlation between treatment effect estimate bias and percentage of covariates with pooled absolute standardized bias >.1 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average ICC	Cluster size	RIS model P	RIS model 2S	RIS model WC	OP model P	OP model 2S	OP model WC	RI model P	RI model 2S	RI model WC	SL model P	SL model 2S	SL model WC	NoL2 model P	NoL2 model 2S	NoL2 model WC	No match
.08	10	0.15	0.06	0.24	0.09	0.10	0.27	0.22	0.29	0.26	-0.05	-0.08	0.32	0.10	0.03	0.32	-0.09
	25	0.40	0.42	0.45	0.40	0.34	0.42	0.46	0.49	0.30	0.07	0.01	0.42	0.13	0.04	0.40	-0.08
	100	0.28	0.26	0.01	0.19	0.24	-0.09	0.02	0.12	0.02			0.09	0.03	0.08	0.07	0.01
	400														0.07		
.15	10	0.08	0.08	0.29	0.10	0.10	0.24	0.34	0.28	0.27	-0.02	-0.04	0.35	0.04	0.13	0.35	-0.01
	25	0.43	0.42	0.38	0.37	0.41	0.35	0.40	0.44	0.29	0.02	-0.02	0.34	0.02	-0.01	0.31	0.02
	100	0.17	0.27	0.00	0.19	0.11	-0.02			0.09			0.12		-0.03	0.06	
	400																
.27	10	0.16	0.14	0.34	0.10	0.10	0.33	0.29	0.22	0.28	0.05	-0.05	0.34	0.04	0.03	0.36	0.02
	25	0.41	0.34	0.39	0.41	0.40	0.45	0.43	0.54	0.34	-0.01	0.08	0.33	0.05	0.07	0.31	
	100	0.10	0.16	0.13	0.15	0.11	-0.01	0.06	0.02	0.10			0.11			0.08	
	400																
.42	10	0.03	0.16	0.27	0.11	0.16	0.29	0.31	0.33	0.31	-0.01	-0.05	0.29	0.09	-0.02	0.31	
	25	0.40	0.40	0.38	0.41	0.42	0.40	0.39	0.46	0.31	-0.03	-0.08	0.19		0.03	0.25	
	100	0.15	0.19	0.06	0.22	0.28	0.09	-0.04	0.07	0.05			0.17			0.20	
	400																

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching. Blank cells are those in which a correlation could not be calculated because all covariates in all replications had ASB<.1.

Table A54.

Correlation between treatment effect estimate bias and percentage of covariates with pooled absolute standardized bias >.1 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method														
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.35	0.31	0.40	0.26	0.32	0.42	0.43	0.46	0.28	-0.01	0.02	0.40	0.06	0.11	0.35	-0.09
	25	0.56	0.61	0.67	0.57	0.56	0.66	0.58	0.61	0.30	0.07	0.01	0.38	0.13	0.05	0.41	-0.08
	100	0.31	0.29	0.13	0.26	0.33	0.10	0.02	0.12	0.02			0.09	0.03	0.08	0.07	0.01
	400														0.07		
.15	10	0.30	0.33	0.41	0.27	0.32	0.39	0.49	0.46	0.32	-0.08	-0.02	0.40	0.05	0.14	0.37	-0.01
	25	0.62	0.59	0.65	0.58	0.57	0.61	0.52	0.58	0.31	0.02	-0.02	0.31	0.01	-0.01	0.29	0.02
	100	0.23	0.30	0.12	0.24	0.16	0.10			0.09			0.12		-0.03	0.06	
	400																
.27	10	0.35	0.34	0.44	0.34	0.40	0.43	0.46	0.46	0.33	0.11	-0.03	0.37	-0.03	0.12	0.40	0.02
	25	0.58	0.57	0.63	0.65	0.64	0.62	0.58	0.62	0.29	-0.01	0.08	0.23	0.05	0.06	0.22	
	100	0.17	0.24	0.24	0.22	0.20	0.17	0.12	0.02	0.10			0.11			0.08	
	400																
.42	10	0.28	0.39	0.41	0.35	0.41	0.41	0.47	0.54	0.37	-0.03	0.02	0.31	0.12	-0.01	0.34	
	25	0.66	0.56	0.63	0.61	0.62	0.64	0.52	0.54	0.30	-0.03	-0.08	0.08		0.03	0.19	
	100	0.25	0.26	0.21	0.24	0.32	0.21	0.03	0.07	0.05			0.17			0.20	
	400																

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching. Blank cells are those in which a correlation could not be calculated because all covariates in all replications had ASB<.1.

Table A55.

Correlation between treatment effect estimate bias and percentage of covariates with pooled absolute standardized bias >.25 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average ICC	Cluster size	RIS model P	RIS model 2S	RIS model WC	OP model P	OP model 2S	OP model WC	RI model P	RI model 2S	RI model WC	SL model P	SL model 2S	SL model WC	NoL2 model P	NoL2 model 2S	NoL2 model WC	No match
.08	10	0.39	0.38	0.42	0.40	0.43	0.42	0.47	0.40	0.41			0.47	0.06	0.04	0.42	-0.10
	25	0.41	0.34	0.22	0.40	0.39	0.24	-0.06		0.28			0.12	-0.03	0.00	0.20	-0.01
	100																
	400																
.15	10	0.40	0.41	0.40	0.44	0.40	0.35	0.58	0.49	0.33			0.43	0.13	0.15	0.47	0.00
	25	0.38	0.39	0.23	0.41	0.31	0.29	0.06		0.05			0.14	0.09	0.04	0.07	-0.04
	100													0.17	0.03		-0.02
	400													0.12			
.27	10	0.42	0.48	0.54	0.42	0.47	0.50	0.45	0.35	0.39			0.40	0.01	-0.01	0.45	-0.05
	25	0.41	0.41	0.20	0.44	0.39	0.30	0.07	0.15	0.10			0.17	0.11	0.09	0.26	-0.03
	100													0.08	0.07		-0.04
	400														0.10		-0.02
.42	10	0.42	0.41	0.44	0.45	0.45	0.49	0.47	0.50	0.39	0.01		0.44	0.07	0.09	0.44	0.02
	25	0.32	0.40	0.32	0.36	0.38	0.35	0.01		0.15			0.03		0.01	0.11	
	100																
	400																

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching. Blank cells are those in which a correlation could not be calculated because all covariates in all replications had ASB<.25.

Table A56.

Correlation between treatment effect estimate bias and percentage of covariates with pooled absolute standardized bias >.25 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method														
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.54	0.54	0.59	0.55	0.56	0.61	0.55	0.49	0.45			0.51	0.06	0.04	0.48	-0.10
	25	0.44	0.39	0.39	0.42	0.45	0.42	-0.06		0.30			0.12	-0.03	0.00	0.20	-0.01
	100																
	400																
.15	10	0.56	0.55	0.57	0.57	0.54	0.59	0.65	0.54	0.46			0.47	0.13	0.15	0.50	0.00
	25	0.39	0.44	0.42	0.44	0.36	0.43	0.06		0.05			0.14	0.09	0.04	0.07	-0.04
	100													0.17	0.03		-0.02
	400													0.12			
.27	10	0.53	0.58	0.66	0.59	0.56	0.64	0.54	0.47	0.44			0.46	0.01	-0.01	0.45	-0.05
	25	0.43	0.43	0.39	0.46	0.44	0.47	0.07	0.16	0.10			0.17	0.11	0.09	0.26	-0.03
	100													0.08	0.07		-0.04
	400														0.10		-0.02
.42	10	0.56	0.57	0.64	0.59	0.57	0.64	0.57	0.58	0.40	0.01		0.42	0.07	0.09	0.46	0.01
	25	0.36	0.42	0.43	0.36	0.43	0.49	0.02		0.15			0.03		0.01	0.11	
	100																
	400																

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching. Blank cells are those in which a correlation could not be calculated because all covariates in all replications had ASB<.25.

Table A57.

Correlation between treatment effect estimate bias and pooled variance ratio with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.01	0.05	0.07	0.05	0.00	0.06	-0.03	-0.05	0.06	0.03	-0.02	0.06	0.00	-0.02	-0.01	-0.05
	25	-0.09	0.05	0.05	0.00	-0.01	0.05	0.06	-0.03	0.10	0.05	0.02	0.03	-0.04	0.02	0.02	-0.04
	100	0.08	0.00	0.06	-0.02	-0.04	0.02	0.02	-0.12	-0.04	-0.01	-0.12	-0.03	-0.05	-0.03	-0.04	0.01
	400	-0.04	-0.03	0.01	-0.02	-0.10	-0.10	-0.10	-0.07	-0.05	0.03	-0.13	0.00	0.01	0.00	0.01	-0.03
.15	10	-0.02	-0.01	0.09	0.03	-0.01	0.10	0.03	-0.01	0.07	-0.02	-0.01	0.04	-0.02	0.04	0.04	-0.02
	25	-0.01	-0.06	-0.04	-0.06	-0.06	-0.06	-0.03	-0.04	-0.01	0.09	0.03	0.03	0.03	0.00	0.00	0.01
	100	0.07	-0.01	0.00	0.11	0.02	0.07	-0.06	-0.05	-0.03	0.05	-0.08	0.01	0.04	0.00	-0.05	-0.03
	400	0.02	-0.07	-0.09	-0.01	0.08	0.06	-0.09	-0.04	-0.04	-0.06	0.01	-0.06	0.03	-0.03	-0.07	-0.03
.27	10	0.00	0.02	0.02	0.02	0.02	0.06	0.01	0.02	0.08	-0.09	-0.06	0.09	-0.03	-0.10	0.09	-0.07
	25	0.04	-0.01	-0.11	-0.04	-0.01	-0.03	-0.08	-0.06	-0.01	-0.02	-0.04	-0.02	-0.04	-0.05	-0.01	-0.03
	100	0.00	0.03	0.01	0.05	0.06	0.03	-0.08	-0.07	-0.02	0.00	-0.10	-0.03	0.07	-0.03	-0.07	0.02
	400	-0.02	0.01	-0.01	0.03	-0.04	0.04	-0.01	0.00	0.03	-0.08	-0.10	0.11	-0.07	-0.02	0.07	0.03
.42	10	-0.03	0.01	-0.01	-0.05	-0.01	0.04	-0.01	-0.02	-0.03	0.00	-0.01	0.01	0.03	-0.03	0.00	-0.01
	25	-0.05	0.08	0.05	0.00	-0.02	0.07	0.03	-0.12	0.01	-0.07	-0.07	-0.06	0.00	0.05	-0.02	0.04
	100	0.06	0.04	0.03	-0.02	0.03	-0.01	-0.16	-0.04	-0.03	0.03	0.02	0.09	-0.05	-0.04	0.07	0.05
	400	0.07	0.02	-0.03	-0.02	0.02	-0.01	0.03	-0.12	-0.14	0.03	-0.02	-0.02	0.02	0.03	-0.06	-0.03

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A58.

Correlation between treatment effect estimate bias and pooled variance ratio with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.03	0.02	0.05	0.02	0.00	0.00	-0.07	-0.04	0.06	0.03	-0.02	0.06	-0.04	-0.04	0.02	-0.06
	25	-0.11	0.03	0.05	-0.04	-0.05	0.08	0.04	0.01	0.08	0.02	-0.03	0.01	-0.06	-0.03	0.01	0.00
	100	0.09	0.03	0.03	0.00	0.00	0.00	0.05	-0.12	-0.07	0.01	0.00	-0.05	0.00	-0.03	0.00	0.06
	400	-0.03	-0.06	0.00	0.01	-0.08	-0.08	-0.07	-0.11	-0.08	-0.01	-0.03	-0.01	0.05	0.12	-0.03	-0.04
.15	10	-0.02	0.01	0.13	0.03	0.04	0.11	0.02	0.01	0.10	-0.02	-0.02	0.05	-0.01	-0.01	0.07	0.04
	25	0.02	-0.01	-0.04	-0.03	-0.07	0.01	-0.03	-0.02	-0.02	0.10	-0.02	0.05	0.04	0.03	0.00	0.03
	100	0.06	-0.03	0.00	0.11	0.03	0.06	-0.02	0.00	0.02	0.02	-0.13	0.04	-0.01	-0.02	-0.02	0.00
	400	0.00	-0.05	-0.06	-0.03	0.04	0.05	-0.07	-0.06	-0.02	0.01	0.00	-0.04	0.04	-0.01	-0.03	-0.02
.27	10	-0.03	0.07	0.02	0.03	0.01	0.03	0.01	0.00	0.06	-0.07	-0.03	0.10	-0.03	-0.11	0.10	-0.01
	25	0.05	-0.01	-0.10	-0.03	-0.03	-0.03	-0.11	-0.06	0.02	0.00	0.00	0.00	0.02	0.03	0.01	0.00
	100	0.04	0.02	-0.02	0.07	0.06	0.01	-0.04	-0.08	0.00	0.00	-0.06	0.02	0.02	-0.03	0.01	0.03
	400	-0.07	0.01	0.00	0.03	-0.03	0.05	-0.02	-0.03	0.03	-0.10	-0.06	0.06	-0.05	0.02	0.01	0.03
.42	10	-0.03	0.02	0.03	-0.06	-0.02	0.05	0.02	0.01	-0.02	0.01	0.01	0.02	0.02	0.02	-0.01	0.04
	25	-0.08	0.05	0.04	-0.09	-0.05	0.06	0.00	-0.07	0.00	-0.09	-0.08	-0.03	-0.03	0.01	-0.09	0.07
	100	0.06	0.02	0.03	-0.06	-0.02	-0.01	-0.14	-0.06	-0.06	0.02	0.02	0.10	-0.02	-0.01	0.07	0.10
	400	0.09	0.08	0.01	0.00	0.06	0.01	0.03	-0.11	-0.14	0.04	0.05	-0.06	0.04	0.02	-0.08	0.01

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A59.

Correlation between treatment effect estimate bias and percentage of covariates with pooled variance ratio <.5 or >2 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method														
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	-0.01	0.01	0.13	0.03	-0.02	0.08	-0.02	0.04	-0.01			0.07			-0.03	
	25	-0.01					0.06										
	100																
	400																
.15	10	0.05	-0.04	0.10	0.04	0.10	0.07	-0.01	-0.03	0.07			0.00			0.01	
	25	0.03		0.03			0.03	-0.04									
	100																
	400																
.27	10	-0.03	0.11	-0.06	-0.02	0.08	0.09	0.00	0.01	0.06	0.00		0.04			0.05	
	25			0.01			0.00			0.05							
	100																
	400																
.42	10	0.03	0.03	0.05	0.01	0.05	0.05	0.02	-0.07	-0.04			-0.05			-0.02	
	25			0.00			0.10										
	100																
	400																

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching. Blank cells are those in which a correlation could not be calculated because all covariates in all replications had VR between .5 and 2.

Table A60.

Correlation between treatment effect estimate bias and percentage of covariates with pooled variance ratio <.5 or >2 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method														
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.04	0.01	0.11	0.00	-0.01	0.03	-0.02	-0.02	0.01			0.08			0.01	
	25	-0.01					0.06										
	100																
	400																
.15	10	0.01	-0.03	0.13	0.07	0.12	0.14	-0.01	-0.01	0.12			0.04			0.04	
	25	0.03		0.03			0.04	-0.04									
	100																
	400																
.27	10	-0.03	0.06	-0.02	-0.01	0.06	0.04	0.02	0.00	0.04	0.00		0.03			0.01	
	25			0.01			0.00			0.05							
	100																
	400																
.42	10	0.04	-0.01	0.08	-0.02	0.05	0.02	0.03	-0.06	-0.03			-0.03			-0.04	
	25			0.00			0.10										
	100																
	400																

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching. Blank cells are those in which a correlation could not be calculated because all covariates in all replications had VR between .5 and 2.

Table A61.

Correlation between treatment effect estimate bias and within-cluster absolute standardized bias (mean across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.10	0.06	0.07	0.06	0.09	0.03	0.06	0.03	0.04	-0.03	0.03	-0.03	0.03	-0.01	0.03	0.42
	25	0.13	0.11	0.09	0.13	0.04	0.04	0.00	0.04	0.12	0.12	0.07	-0.02	0.15	0.12	0.02	0.57
	100	0.00	0.00	0.00	0.01	0.03	0.04	-0.14	-0.04	-0.10	0.13	0.09	-0.11	0.06	0.07	-0.08	0.45
	400	-0.03	-0.04	0.00	0.05	0.01	0.01	-0.12	-0.07	-0.09	0.07	0.11	-0.10	0.00	0.02	-0.06	0.45
.15	10	0.01	0.07	0.09	0.02	0.11	0.05	-0.02	0.01	-0.03	0.04	0.08	0.15	0.09	0.09	0.01	0.48
	25	0.12	0.01	0.01	0.10	0.10	0.03	0.06	0.01	-0.06	0.07	0.04	0.05	0.10	0.10	0.08	0.51
	100	-0.01	0.07	0.00	-0.04	0.09	0.06	0.00	-0.02	0.05	0.11	0.05	-0.03	0.05	0.00	-0.02	0.48
	400	0.05	0.06	0.02	0.04	0.10	0.05	-0.08	0.03	0.06	0.06	0.06	-0.01	-0.01	0.04	0.00	0.48
.27	10	0.07	0.04	0.16	0.07	0.08	0.04	0.10	0.04	0.02	-0.03	0.02	-0.02	0.07	0.04	-0.02	0.47
	25	0.14	0.12	0.03	0.07	0.07	0.07	-0.03	0.05	-0.01	0.11	0.03	0.04	0.01	0.07	-0.01	0.50
	100	0.01	0.01	0.02	0.05	0.07	-0.02	-0.11	-0.08	-0.11	-0.03	0.02	-0.08	0.05	0.08	-0.09	0.43
	400	-0.04	0.09	0.10	-0.09	0.16	0.11	-0.06	0.07	0.06	0.10	0.06	0.06	0.04	0.06	0.07	0.49
.42	10	0.08	0.09	0.09	0.02	0.09	0.15	0.05	0.02	0.00	0.05	0.09	0.01	-0.02	-0.06	-0.01	0.48
	25	0.07	0.15	0.08	0.11	0.12	0.07	0.05	0.05	-0.02	0.13	0.08	0.04	0.14	0.13	-0.06	0.51
	100	0.02	-0.06	0.00	0.00	0.02	0.03	-0.12	-0.01	-0.03	0.02	0.08	-0.08	0.00	0.04	-0.12	0.51
	400	0.07	0.03	0.01	0.01	0.00	-0.01	-0.17	-0.02	-0.01	0.10	0.15	-0.05	0.05	0.21	-0.10	0.51

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A62.

Correlation between treatment effect estimate bias and within-cluster absolute standardized bias (mean across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.10	0.08	0.07	0.06	0.14	0.03	0.09	0.08	0.03	-0.03	0.01	0.00	0.03	0.02	0.05	0.49
	25	0.14	0.14	0.09	0.18	0.11	0.10	-0.05	0.04	0.17	0.14	0.09	-0.02	0.16	0.13	0.04	0.62
	100	0.07	0.07	0.04	0.04	0.11	0.10	-0.14	-0.05	-0.08	0.13	0.11	-0.11	0.08	0.12	-0.08	0.54
	400	-0.02	-0.03	0.02	0.11	-0.02	-0.02	-0.15	0.01	-0.04	0.13	0.16	-0.07	0.03	0.08	-0.06	0.51
.15	10	0.05	0.09	0.05	0.03	0.16	0.06	0.00	0.03	-0.03	0.05	0.04	0.16	0.09	0.09	0.02	0.50
	25	0.21	0.11	0.04	0.16	0.18	0.04	0.02	0.01	-0.12	0.11	0.07	0.03	0.11	0.12	0.05	0.56
	100	-0.01	0.10	0.09	0.01	0.14	0.09	0.00	-0.01	0.05	0.20	0.11	-0.02	0.13	0.09	-0.04	0.56
	400	0.04	0.01	0.03	0.02	0.12	0.04	-0.10	0.04	0.08	0.09	0.13	-0.01	0.02	0.11	0.01	0.57
.27	10	0.08	0.07	0.20	0.13	0.09	0.08	0.06	0.03	0.05	0.01	0.03	0.04	0.07	0.07	0.10	0.53
	25	0.18	0.13	0.08	0.14	0.09	0.09	-0.02	0.04	-0.05	0.12	0.08	0.01	0.01	0.07	-0.04	0.58
	100	0.05	0.05	0.02	0.04	0.08	-0.04	-0.13	-0.07	-0.09	0.02	0.03	-0.08	0.06	0.07	-0.06	0.47
	400	-0.03	0.13	0.14	-0.05	0.12	0.09	-0.09	0.04	0.08	0.14	0.11	0.07	0.10	0.13	0.05	0.54
.42	10	0.11	0.14	0.08	0.09	0.14	0.19	0.07	0.02	0.01	0.03	0.06	0.06	0.01	-0.06	0.03	0.53
	25	0.12	0.18	0.14	0.17	0.14	0.11	0.00	0.00	-0.05	0.13	0.10	0.00	0.15	0.14	-0.10	0.57
	100	0.11	0.02	0.03	0.03	0.07	0.01	-0.12	0.02	-0.05	0.07	0.16	-0.10	0.04	0.11	-0.13	0.57
	400	0.09	0.06	0.02	0.06	0.04	0.03	-0.19	0.01	0.00	0.13	0.22	-0.11	0.10	0.26	-0.11	0.59

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A63.

Correlation between treatment effect estimate bias and within-cluster absolute standardized bias (median across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
		RIS model			OP model			RI model			SL model			NoL2 model			No
Average ICC	Cluster size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.09	0.16	0.09	0.11	0.16	0.08	0.00	0.04	0.05	0.07	0.01	0.00	0.05	-0.01	0.01	0.41
	25	0.24	0.16	0.09	0.14	0.16	0.06	0.06	0.04	0.02	0.06	0.13	-0.05	0.12	0.06	0.03	0.51
	100	0.11	-0.05	0.00	0.11	-0.04	-0.04	-0.11	-0.07	-0.03	0.09	0.03	-0.05	0.08	0.00	-0.03	0.37
	400	0.02	-0.01	0.03	0.09	-0.04	-0.01	-0.04	-0.01	-0.02	0.09	0.05	0.01	-0.04	0.04	0.02	0.32
.15	10	0.02	0.08	0.13	0.04	0.06	0.07	0.02	0.07	0.05	0.05	0.09	0.16	0.08	0.08	0.05	0.48
	25	0.16	0.09	0.14	0.18	0.06	0.08	0.05	0.03	0.01	0.10	0.07	0.03	0.06	0.10	0.07	0.47
	100	0.05	0.07	0.06	0.03	0.08	0.08	-0.02	0.04	0.01	0.12	0.06	-0.05	0.09	0.06	0.02	0.39
	400	0.03	0.05	0.00	0.07	0.09	0.13	-0.06	0.00	0.06	0.06	0.04	0.04	-0.03	0.02	0.07	0.32
.27	10	0.16	0.06	0.18	0.05	0.07	0.10	0.11	0.02	0.04	0.01	0.05	0.04	0.04	0.01	0.07	0.42
	25	0.13	0.13	0.04	0.14	0.15	0.05	-0.02	0.11	0.00	0.12	0.08	0.04	0.04	0.02	0.07	0.46
	100	0.13	0.04	0.01	0.12	0.03	-0.02	-0.10	-0.04	0.03	-0.09	-0.02	0.01	0.02	-0.03	-0.01	0.40
	400	0.03	0.02	0.01	0.00	0.07	0.13	-0.08	0.04	0.07	0.05	0.03	0.04	-0.02	-0.04	0.06	0.39
.42	10	0.03	0.02	0.11	-0.01	0.15	0.18	-0.01	0.05	-0.04	0.01	0.02	0.01	0.06	0.01	0.02	0.45
	25	0.13	0.21	0.13	0.20	0.19	0.16	-0.04	0.09	0.03	0.13	0.13	0.10	0.12	0.13	0.04	0.48
	100	0.06	0.09	0.10	0.10	0.13	0.02	-0.07	-0.01	0.03	-0.01	0.05	-0.06	0.03	0.04	-0.06	0.41
	400	0.09	0.13	0.04	0.00	0.01	-0.02	-0.12	-0.09	0.01	-0.01	0.14	0.02	0.10	0.19	-0.05	0.29

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A64.

Correlation between treatment effect estimate bias and within-cluster absolute standardized bias (median across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.10	0.18	0.10	0.09	0.20	0.10	0.03	0.04	0.03	0.07	0.00	0.03	0.06	-0.03	0.03	0.44
	25	0.25	0.23	0.13	0.18	0.20	0.15	0.00	0.03	0.03	0.09	0.15	-0.01	0.12	0.07	0.08	0.51
	100	0.19	0.04	0.06	0.17	0.06	0.04	-0.10	-0.05	0.02	0.10	0.06	-0.01	0.10	0.06	-0.01	0.39
	400	0.02	0.03	0.07	0.14	-0.02	-0.03	-0.08	0.03	0.02	0.13	0.08	0.02	-0.01	0.06	0.03	0.29
.15	10	0.08	0.10	0.07	0.05	0.11	0.08	0.08	0.08	0.03	0.06	0.11	0.16	0.08	0.07	0.07	0.48
	25	0.26	0.17	0.20	0.25	0.17	0.19	0.00	0.02	0.00	0.11	0.10	0.04	0.09	0.10	0.07	0.46
	100	0.04	0.13	0.14	0.07	0.11	0.13	-0.02	0.07	0.05	0.16	0.06	-0.06	0.13	0.11	0.01	0.40
	400	0.01	-0.02	-0.03	0.06	0.03	0.06	-0.09	-0.03	0.05	0.06	0.06	0.05	-0.03	0.07	0.10	0.34
.27	10	0.19	0.10	0.21	0.09	0.13	0.14	0.06	0.01	0.06	0.06	0.05	0.03	0.05	0.01	0.08	0.44
	25	0.22	0.19	0.09	0.21	0.27	0.09	-0.06	0.10	-0.03	0.14	0.11	0.03	0.06	0.04	0.04	0.49
	100	0.15	0.13	0.06	0.15	0.09	0.02	-0.10	-0.05	0.02	-0.02	-0.03	0.03	0.01	-0.04	-0.03	0.40
	400	0.01	0.06	0.07	0.06	0.05	0.10	-0.10	0.04	0.07	0.05	0.04	0.03	0.03	-0.01	0.02	0.37
.42	10	0.11	0.05	0.13	0.06	0.15	0.21	0.01	0.02	-0.06	0.00	0.02	0.04	0.08	0.04	0.06	0.50
	25	0.16	0.24	0.18	0.26	0.25	0.20	-0.03	0.06	0.03	0.13	0.14	0.08	0.13	0.13	0.02	0.52
	100	0.14	0.10	0.17	0.11	0.19	0.00	-0.09	-0.03	0.02	0.02	0.09	-0.10	0.06	0.06	-0.09	0.41
	400	0.09	0.12	0.00	0.06	0.02	0.01	-0.13	-0.08	0.02	0.00	0.19	0.00	0.11	0.21	-0.07	0.31

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A65.

Correlation between treatment effect estimate bias and percentage of clusters with pooled absolute standardized bias >.1 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.08	0.11	0.06	0.03	0.02	0.01	-0.02	0.04	0.01	-0.04	0.02	-0.07	-0.07	0.00	-0.04	0.09
	25	0.15	0.15	0.07	0.10	0.09	0.11	-0.04	0.00	0.00	0.04	0.03	0.04	0.03	0.04	-0.01	0.13
	100	0.10	0.00	0.02	0.11	0.04	0.02	-0.02	0.00	-0.10	0.05	0.03	0.03	-0.01	0.04	0.03	0.11
	400	-0.01	0.00	0.06	0.11	-0.02	-0.01	0.00	0.02	-0.04	-0.07	-0.02	-0.08	-0.07	0.02	0.02	0.11
.15	10	0.06	0.00	-0.06	-0.02	-0.06	0.05	0.02	-0.01	0.00	0.02	0.00	0.05	-0.03	-0.03	0.00	0.10
	25	0.13	0.10	0.02	0.06	0.02	0.03	-0.02	0.05	0.03	-0.01	0.02	-0.03	0.06	0.02	0.00	0.09
	100	0.08	0.06	0.05	0.04	0.00	0.11	0.01	0.05	0.00	-0.02	-0.07	0.03	0.01	-0.06	0.05	0.14
	400	0.03	0.02	0.01	0.05	0.11	0.14	0.05	0.03	0.07	-0.05	-0.10	0.09	-0.06	-0.06	0.00	0.11
.27	10	0.08	-0.07	0.04	0.07	0.09	-0.07	0.03	-0.05	0.00	-0.04	0.02	-0.02	0.00	0.05	-0.01	0.11
	25	0.12	0.11	0.05	0.10	0.12	0.16	0.03	0.07	0.02	0.06	-0.03	0.03	0.03	-0.05	0.04	0.10
	100	0.08	0.05	0.07	0.08	0.07	-0.03	-0.04	-0.05	0.06	-0.05	0.03	0.03	0.07	0.08	0.01	0.10
	400	0.07	0.01	0.03	-0.02	0.12	0.15	0.02	-0.05	0.05	0.01	0.00	0.12	0.02	-0.02	0.04	0.18
.42	10	0.00	0.05	0.03	0.02	0.13	0.11	0.07	0.03	-0.02	-0.01	0.01	-0.09	-0.01	0.03	0.01	0.08
	25	0.08	-0.02	0.09	0.13	0.16	0.16	0.00	0.05	0.10	0.03	0.08	-0.01	0.05	0.07	0.02	0.12
	100	0.07	0.04	0.04	0.12	0.10	-0.02	-0.06	-0.01	-0.03	0.02	-0.04	-0.07	0.01	0.04	-0.06	0.20
	400	0.04	0.10	0.04	0.01	0.05	0.01	-0.05	-0.07	0.07	-0.06	-0.01	0.05	-0.05	0.03	-0.01	0.16

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A66.

Correlation between treatment effect estimate bias and percentage of clusters with pooled absolute standardized bias >.1 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.04	0.05	0.07	0.08	0.02	0.06	0.02	0.08	-0.03	-0.04	0.02	-0.03	-0.11	-0.04	-0.02	0.14
	25	0.18	0.18	0.08	0.14	0.12	0.14	-0.04	0.01	-0.02	0.05	0.05	0.01	0.02	0.13	-0.01	0.09
	100	0.16	0.08	0.05	0.11	0.08	0.02	0.02	0.00	-0.06	0.10	0.02	0.02	0.00	0.07	0.04	0.09
	400	0.02	0.01	0.08	0.14	-0.03	0.00	-0.01	0.06	-0.02	-0.04	-0.05	-0.01	-0.04	0.06	0.05	0.13
.15	10	0.05	0.00	0.00	0.02	-0.01	0.06	0.03	0.04	0.02	-0.03	0.01	0.03	-0.02	0.03	0.04	0.11
	25	0.12	0.14	0.03	0.10	0.04	0.07	0.01	0.02	0.05	-0.03	0.02	0.00	0.03	0.01	0.02	0.10
	100	0.08	0.10	0.09	0.06	0.02	0.12	0.01	0.09	0.06	0.09	-0.01	0.05	0.09	0.02	0.07	0.12
	400	0.04	-0.05	0.00	0.05	0.06	0.06	0.02	0.03	0.10	-0.02	-0.08	0.13	-0.01	0.00	0.06	0.09
.27	10	0.11	0.01	0.05	0.09	0.14	-0.01	0.01	-0.02	-0.01	-0.01	0.03	-0.01	0.03	0.10	-0.01	0.18
	25	0.16	0.13	0.10	0.13	0.16	0.18	-0.01	0.07	0.06	0.02	0.02	0.00	-0.02	0.03	0.02	0.15
	100	0.11	0.09	0.07	0.08	0.10	0.02	-0.05	-0.05	0.03	0.03	0.06	0.06	0.04	0.12	0.01	0.08
	400	0.03	0.05	0.08	0.05	0.10	0.11	0.05	-0.03	0.07	0.06	-0.03	0.10	0.06	-0.04	0.02	0.17
.42	10	0.01	0.04	0.08	0.04	0.11	0.11	0.05	0.02	-0.01	0.05	0.03	-0.02	0.00	0.02	0.04	0.14
	25	0.07	0.04	0.13	0.11	0.16	0.17	0.00	0.03	0.05	0.03	0.02	-0.02	0.05	0.11	0.01	0.13
	100	0.10	0.07	0.09	0.09	0.13	-0.05	-0.03	0.00	-0.07	0.03	0.00	-0.03	-0.01	0.05	-0.06	0.17
	400	0.04	0.10	0.03	0.07	0.07	0.04	0.02	-0.02	0.12	0.00	0.02	0.05	0.02	0.09	0.01	0.15

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A67.

Correlation between treatment effect estimate bias and percentage of clusters with pooled absolute standardized bias >.25 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.08	0.14	0.03	0.12	0.11	0.02	-0.01	0.08	0.04	-0.01	0.01	0.04	0.02	0.02	0.06	0.19
	25	0.21	0.15	0.05	0.19	0.21	0.08	0.03	0.10	0.03	0.07	0.07	0.01	0.04	0.04	-0.02	0.17
	100	0.15	0.01	0.03	0.14	0.03	0.05	-0.11	-0.07	-0.04	0.02	0.00	-0.06	-0.02	-0.01	-0.02	0.11
	400	-0.03	-0.06	-0.05	0.05	-0.03	-0.01	-0.09	-0.06	-0.03	-0.02	-0.01	0.07	0.01	0.07	0.02	0.13
.15	10	0.07	0.08	-0.02	0.05	-0.01	0.01	0.00	0.00	-0.02	0.00	0.00	0.07	0.02	0.05	0.00	0.18
	25	0.19	0.12	0.12	0.17	0.04	0.04	-0.04	0.08	0.00	0.04	0.01	-0.02	0.05	0.04	0.03	0.14
	100	0.07	0.12	0.09	0.07	0.12	0.13	0.02	0.05	0.04	0.05	0.03	0.06	0.00	-0.01	-0.03	0.09
	400	0.06	0.05	0.03	0.09	0.08	0.00	0.00	0.03	0.05	0.00	-0.04	0.01	-0.06	-0.07	0.00	0.15
.27	10	0.12	0.00	0.08	0.12	0.06	0.07	0.00	-0.01	-0.08	-0.06	0.02	0.02	0.02	0.04	0.03	0.18
	25	0.13	0.16	0.06	0.13	0.16	0.08	0.04	0.02	0.02	0.06	0.01	-0.01	0.01	-0.02	0.00	0.09
	100	0.12	0.07	0.07	0.13	0.04	-0.01	-0.03	-0.08	0.09	-0.11	0.01	0.03	0.00	0.11	0.06	0.15
	400	0.00	0.11	0.11	-0.10	0.12	0.05	-0.05	0.01	0.09	0.04	-0.05	0.09	-0.01	-0.06	0.11	0.20
.42	10	0.05	0.05	0.08	0.06	0.10	0.14	0.00	0.12	-0.03	0.07	0.03	-0.07	0.03	-0.02	-0.06	0.18
	25	0.19	0.17	0.07	0.20	0.20	0.14	-0.03	0.07	0.06	0.01	0.09	0.00	0.09	0.08	-0.02	0.18
	100	0.08	0.13	0.13	0.08	0.12	0.06	-0.04	0.01	-0.04	-0.02	-0.02	-0.05	-0.02	0.07	-0.04	0.18
	400	0.05	0.04	0.01	0.00	-0.02	-0.01	-0.10	-0.06	-0.04	-0.02	-0.01	0.00	-0.05	0.07	0.00	0.25

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A68.

Correlation between treatment effect estimate bias and percentage of clusters with pooled absolute standardized bias >.25 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.08	0.13	0.07	0.12	0.12	0.08	-0.02	0.08	0.00	-0.01	0.03	0.07	0.03	-0.02	0.08	0.25
	25	0.26	0.18	0.08	0.22	0.22	0.15	0.00	0.10	0.01	0.08	0.08	-0.01	0.03	0.11	0.00	0.14
	100	0.20	0.11	0.11	0.16	0.12	0.14	-0.11	-0.04	0.03	0.09	0.07	-0.05	0.01	0.05	0.00	0.13
	400	-0.02	-0.06	-0.05	0.10	0.00	0.03	-0.10	0.01	0.03	0.05	0.04	0.08	0.06	0.07	0.06	0.12
.15	10	0.09	0.07	0.02	0.08	0.07	0.07	0.00	0.06	0.02	-0.03	0.07	0.08	0.05	0.07	0.06	0.24
	25	0.24	0.14	0.15	0.19	0.09	0.14	-0.01	0.09	0.04	0.04	0.00	-0.02	0.06	0.07	0.05	0.16
	100	0.07	0.18	0.17	0.10	0.16	0.17	0.04	0.08	0.06	0.15	0.12	0.07	0.13	0.08	-0.03	0.11
	400	0.02	0.03	0.07	0.03	0.10	0.00	-0.05	0.04	0.05	0.02	-0.01	0.04	0.02	0.07	0.02	0.12
.27	10	0.15	0.05	0.08	0.16	0.13	0.08	-0.01	0.01	-0.04	-0.01	0.02	0.03	0.06	0.07	0.02	0.25
	25	0.16	0.17	0.08	0.20	0.20	0.12	-0.02	0.05	0.03	0.06	0.06	-0.01	0.00	0.01	0.01	0.23
	100	0.15	0.13	0.09	0.14	0.09	0.03	0.00	-0.07	0.08	-0.01	0.03	0.03	0.02	0.12	0.04	0.17
	400	0.01	0.15	0.10	-0.05	0.09	0.04	-0.08	-0.02	0.08	0.09	0.03	0.10	0.09	0.01	0.10	0.21
.42	10	0.04	0.09	0.08	0.10	0.09	0.16	-0.04	0.06	0.00	0.10	0.05	-0.01	0.06	-0.03	-0.03	0.26
	25	0.22	0.20	0.10	0.23	0.25	0.14	-0.02	0.04	0.03	0.05	0.05	0.03	0.09	0.09	-0.03	0.25
	100	0.14	0.15	0.17	0.10	0.14	0.04	-0.04	0.03	-0.03	0.03	0.10	-0.03	0.02	0.07	-0.04	0.20
	400	0.06	0.04	0.02	0.01	0.01	0.00	-0.07	0.01	0.01	0.06	0.03	0.01	-0.01	0.13	0.03	0.24

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A69.

Correlation between treatment effect estimate bias and within-cluster variance ratio (mean across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.01	0.02	0.12	0.11	0.07	-0.01	0.03	0.02	-0.09	0.01	0.00	0.02	-0.02	-0.01	-0.01	-0.04
	25	0.05	0.10	0.08	0.06	0.11	-0.01	-0.02	0.01	0.04	0.01	0.06	0.05	0.10	-0.04	0.08	-0.02
	100	-0.06	0.03	-0.08	-0.02	0.03	-0.04	-0.03	0.01	-0.05	0.10	0.01	-0.03	-0.03	-0.04	-0.05	-0.06
	400	0.06	0.00	-0.01	0.00	-0.02	-0.07	0.03	-0.04	-0.02	-0.01	0.03	0.04	0.04	0.01	-0.01	0.02
.15	10	0.03	0.06	0.00	-0.02	0.02	-0.06	-0.03	0.05	0.00	-0.02	0.02	0.09	-0.01	0.06	0.01	0.02
	25	-0.03	0.00	0.09	0.01	0.01	0.04	0.07	0.11	-0.01	0.04	0.01	-0.04	0.00	-0.07	-0.06	-0.02
	100	0.05	0.01	-0.10	-0.04	0.01	-0.01	-0.01	0.05	0.00	0.09	-0.02	-0.03	0.01	0.01	-0.07	-0.02
	400	-0.03	-0.03	0.01	-0.04	-0.05	-0.01	-0.09	0.06	0.07	0.05	0.07	-0.03	0.04	0.11	0.02	0.03
.27	10	-0.01	0.11	0.02	0.02	0.10	0.09	0.06	-0.03	0.03	0.04	0.08	0.05	-0.03	0.02	0.05	0.07
	25	-0.01	-0.02	-0.02	-0.02	0.02	0.02	0.05	0.15	0.08	0.04	-0.02	0.08	-0.07	0.00	0.09	-0.04
	100	0.00	0.06	-0.06	0.04	0.03	-0.06	0.03	-0.01	-0.09	-0.05	0.03	-0.06	0.05	-0.02	-0.01	-0.07
	400	-0.04	-0.05	0.00	-0.06	-0.06	-0.04	0.00	-0.01	-0.12	0.07	0.05	-0.01	0.06	0.06	0.00	-0.02
.42	10	0.00	-0.04	0.09	-0.03	-0.02	0.05	0.03	0.02	0.01	-0.04	0.02	0.06	-0.05	-0.03	0.05	0.06
	25	0.04	0.05	0.05	0.08	0.11	0.05	0.05	0.08	-0.01	0.10	0.06	0.01	0.06	0.09	-0.02	0.04
	100	-0.01	-0.01	0.02	0.07	-0.05	-0.05	-0.02	0.01	0.00	0.03	-0.01	0.03	0.01	0.05	0.02	-0.05
	400	-0.04	-0.02	0.01	-0.01	-0.12	-0.09	-0.09	-0.08	-0.02	0.01	-0.01	-0.02	0.02	0.04	-0.01	0.05

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A70.

Correlation between treatment effect estimate bias and within-cluster variance ratio (mean across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.00	-0.02	0.10	0.11	0.06	-0.01	0.01	0.02	-0.05	0.03	-0.04	0.07	0.02	-0.01	0.01	-0.06
	25	0.01	0.06	0.07	0.06	0.08	0.00	0.02	0.02	0.02	-0.01	0.05	0.08	0.03	-0.05	0.10	-0.04
	100	-0.05	0.09	-0.03	0.02	0.03	-0.06	-0.04	-0.04	-0.09	0.04	0.05	-0.07	-0.03	-0.01	-0.08	-0.03
	400	0.06	-0.01	-0.04	0.04	0.01	-0.07	0.03	-0.02	-0.02	-0.01	0.01	0.00	0.03	0.02	-0.03	0.06
.15	10	0.01	0.06	-0.02	0.00	0.07	-0.04	-0.04	0.02	0.01	0.02	0.05	0.03	0.04	0.07	-0.02	0.04
	25	-0.01	0.02	0.13	0.02	0.01	0.04	0.07	0.09	-0.03	0.06	0.01	0.00	0.02	-0.07	-0.06	-0.03
	100	0.01	0.07	-0.05	-0.05	0.05	0.00	0.01	0.04	-0.03	0.09	0.00	-0.08	0.02	0.00	-0.11	0.01
	400	-0.04	-0.06	-0.02	-0.08	-0.05	0.02	-0.05	0.04	0.06	0.04	0.06	-0.03	0.02	0.10	0.01	0.05
.27	10	0.01	0.05	0.09	0.05	0.09	0.12	0.07	-0.01	0.02	-0.03	0.04	0.03	-0.05	-0.03	0.04	0.06
	25	0.04	-0.01	-0.02	-0.04	0.00	0.01	0.05	0.15	0.06	0.04	-0.05	0.05	-0.08	-0.04	0.07	-0.04
	100	-0.02	0.01	-0.02	0.06	0.01	-0.02	0.02	-0.05	-0.09	-0.06	0.05	-0.09	0.01	-0.03	-0.02	-0.04
	400	-0.02	-0.01	0.04	-0.05	-0.04	-0.02	-0.03	0.02	-0.07	0.05	0.08	0.01	0.07	0.08	0.02	0.02
.42	10	-0.02	-0.02	0.09	-0.05	-0.03	0.05	0.05	0.01	0.03	-0.02	-0.01	0.06	-0.06	-0.06	0.07	0.04
	25	-0.02	0.03	0.06	0.07	0.15	0.04	-0.01	0.02	-0.01	0.09	0.07	0.01	0.07	0.11	0.00	0.05
	100	-0.02	0.00	0.01	0.06	-0.08	-0.05	-0.04	0.00	-0.02	-0.02	-0.05	-0.02	-0.03	0.05	-0.03	-0.02
	400	-0.02	0.02	0.03	0.01	-0.13	-0.10	-0.06	-0.07	-0.02	0.00	-0.04	-0.01	-0.02	0.01	0.00	0.04

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A71.

Correlation between treatment effect estimate bias and within-cluster variance ratio (median across clusters) with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	-0.01	0.05	0.13	0.10	0.05	-0.04	-0.01	0.00	-0.09	0.04	0.02	0.02	-0.01	0.03	0.00	0.00
	25	0.02	0.10	0.05	0.02	0.10	-0.01	-0.05	0.02	0.06	0.00	0.06	0.05	0.09	-0.06	0.09	-0.03
	100	-0.04	-0.02	-0.11	0.00	0.02	-0.05	-0.04	0.05	-0.05	0.12	0.05	-0.04	-0.02	-0.01	-0.05	-0.04
	400	0.08	0.00	-0.03	0.02	0.04	-0.07	0.02	-0.01	0.02	0.01	0.01	-0.01	0.00	0.06	-0.02	0.02
.15	10	0.02	0.07	-0.03	-0.07	0.02	-0.07	-0.06	0.04	-0.03	-0.02	0.03	0.05	0.01	0.09	-0.01	-0.04
	25	-0.08	0.02	0.10	0.00	0.01	0.03	0.08	0.09	-0.03	0.08	0.02	-0.04	-0.02	-0.05	-0.08	-0.02
	100	0.03	0.03	-0.07	-0.02	-0.03	-0.03	-0.01	0.01	-0.04	0.07	-0.03	-0.03	0.00	0.04	-0.06	-0.01
	400	-0.02	-0.02	0.01	-0.02	-0.05	-0.04	-0.10	0.07	0.08	0.02	0.03	-0.03	0.00	0.04	0.02	0.00
.27	10	0.01	0.07	0.00	0.04	0.05	0.11	0.02	0.01	0.00	0.04	0.03	0.04	-0.08	0.01	0.03	0.03
	25	0.01	-0.04	0.02	-0.03	0.00	-0.02	0.03	0.13	0.08	0.01	-0.02	0.10	-0.07	-0.03	0.11	-0.01
	100	0.03	0.06	-0.02	0.06	0.03	-0.02	0.04	-0.02	-0.11	-0.04	0.06	-0.06	0.03	-0.03	-0.01	-0.05
	400	0.06	-0.03	-0.02	-0.05	-0.04	-0.04	-0.01	-0.05	-0.09	0.04	0.02	0.00	0.05	0.05	0.01	-0.01
.42	10	-0.01	-0.03	0.11	-0.01	-0.02	0.05	0.03	-0.02	0.02	-0.04	-0.02	0.06	-0.05	-0.04	0.01	0.06
	25	0.05	0.07	0.02	0.08	0.10	0.03	0.03	0.05	-0.06	0.10	0.01	-0.01	0.07	0.11	-0.03	-0.01
	100	0.01	0.03	0.04	0.08	-0.04	0.01	-0.03	0.04	0.04	0.01	0.02	-0.02	-0.04	0.05	0.07	-0.05
	400	-0.07	-0.01	0.01	-0.05	-0.07	-0.06	-0.03	-0.08	-0.06	0.04	0.00	0.06	0.05	0.01	0.00	0.05

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A72.

Correlation between treatment effect estimate bias and within-cluster variance ratio (median across clusters) with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	-0.01	0.02	0.10	0.10	0.07	-0.03	-0.02	0.02	-0.05	0.05	-0.03	0.03	0.04	0.03	0.02	-0.02
	25	0.00	0.06	0.03	0.04	0.06	-0.01	0.01	0.01	0.02	0.00	0.06	0.08	0.03	-0.03	0.09	-0.06
	100	-0.02	0.01	-0.07	0.04	0.01	-0.07	-0.03	-0.01	-0.07	0.04	0.03	-0.09	-0.04	0.02	-0.05	-0.03
	400	0.07	0.00	-0.02	0.07	0.05	-0.05	0.04	-0.02	0.00	0.00	0.00	-0.03	0.01	0.09	-0.04	0.06
.15	10	0.00	0.06	-0.05	-0.04	0.08	-0.04	-0.07	0.01	0.00	0.03	0.07	-0.01	0.05	0.11	-0.05	0.01
	25	-0.06	0.02	0.11	0.03	0.01	0.04	0.07	0.06	-0.04	0.05	0.04	0.01	0.00	-0.07	-0.09	-0.02
	100	0.00	0.05	-0.03	-0.03	0.01	-0.01	0.02	0.02	-0.07	0.06	-0.05	-0.05	0.04	0.01	-0.06	0.03
	400	-0.02	-0.02	-0.02	-0.03	-0.02	0.01	-0.08	0.04	0.06	-0.01	0.05	-0.02	0.00	0.05	0.03	0.03
.27	10	0.04	0.02	0.06	0.07	0.06	0.14	0.04	0.04	-0.01	-0.02	0.00	0.05	-0.12	-0.04	0.05	0.03
	25	0.06	-0.06	-0.02	-0.06	-0.01	-0.04	0.04	0.14	0.08	0.01	-0.04	0.05	-0.08	-0.07	0.10	-0.05
	100	0.03	0.02	0.04	0.07	0.03	0.01	0.06	-0.03	-0.09	-0.05	0.08	-0.06	0.01	-0.04	-0.03	-0.02
	400	0.06	0.00	0.03	-0.05	-0.04	-0.03	-0.07	-0.02	-0.06	0.01	0.05	0.01	0.03	0.07	0.02	0.05
.42	10	-0.02	0.00	0.11	-0.02	-0.02	0.04	0.02	-0.02	0.03	-0.03	-0.07	0.07	-0.05	-0.06	0.04	0.05
	25	0.00	0.07	0.03	0.08	0.11	0.03	-0.01	0.01	-0.06	0.11	0.04	0.00	0.08	0.14	-0.03	0.00
	100	-0.01	0.03	0.03	0.07	-0.04	0.00	-0.02	0.06	0.02	-0.03	0.01	-0.03	-0.08	0.05	0.02	-0.04
	400	-0.04	0.01	0.03	0.00	-0.08	-0.07	0.01	-0.07	-0.07	0.02	0.00	0.05	0.03	-0.03	0.00	0.06

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A73.

Correlation between treatment effect estimate bias and percentage of clusters with pooled variance ratio <.5 or >2 with covariates equally weighted, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.06	0.04	0.14	0.09	0.05	0.01	0.05	-0.01	-0.11	-0.01	0.01	0.03	-0.03	0.02	-0.01	-0.05
	25	0.04	0.10	0.10	0.02	0.14	0.02	-0.03	0.00	-0.01	0.00	0.05	0.08	0.12	-0.05	0.10	-0.02
	100	-0.01	0.05	-0.08	-0.05	-0.01	-0.07	0.02	0.02	-0.04	0.06	0.04	-0.05	0.00	-0.04	-0.02	0.00
	400	0.01	-0.01	-0.02	-0.02	0.00	-0.05	0.07	-0.07	-0.03	-0.02	-0.04	0.05	0.05	0.01	-0.01	0.03
.15	10	0.04	0.05	-0.01	-0.03	-0.02	-0.06	-0.06	0.06	0.01	-0.01	0.01	0.07	-0.02	0.07	0.02	-0.06
	25	-0.03	0.03	0.09	0.00	0.00	0.03	0.04	0.06	-0.01	0.08	0.00	-0.05	-0.04	-0.05	-0.07	-0.02
	100	0.05	0.01	-0.10	-0.03	0.06	0.00	0.02	0.02	0.02	0.12	-0.01	-0.04	0.04	0.02	-0.05	0.01
	400	0.03	-0.04	-0.02	-0.07	-0.04	0.03	-0.11	0.01	0.00	0.06	0.11	-0.03	0.03	0.08	-0.08	0.07
.27	10	0.00	0.04	0.02	-0.04	0.06	0.09	0.01	-0.04	0.05	0.06	0.05	0.03	-0.03	-0.02	0.04	0.04
	25	-0.01	-0.02	-0.04	0.01	0.00	0.02	0.04	0.10	0.09	0.02	-0.02	0.08	-0.09	-0.02	0.11	-0.03
	100	-0.01	0.05	-0.06	0.01	-0.01	-0.07	0.01	0.02	-0.04	-0.05	0.02	-0.04	0.08	0.00	-0.03	-0.07
	400	-0.13	-0.02	0.01	-0.03	-0.03	-0.06	-0.02	0.02	-0.04	0.05	0.02	0.00	0.04	0.05	-0.02	-0.09
.42	10	0.00	-0.01	0.08	-0.01	0.01	0.03	0.02	0.02	0.04	-0.05	0.02	0.07	-0.05	-0.04	0.05	0.04
	25	0.03	0.03	0.02	0.05	0.07	0.07	0.05	0.06	-0.02	0.11	0.03	-0.02	0.05	0.08	0.00	-0.03
	100	-0.05	-0.07	0.01	0.07	-0.02	0.03	-0.01	0.00	-0.02	0.02	-0.06	0.03	0.02	0.02	0.05	0.01
	400	-0.03	-0.04	-0.04	0.03	-0.15	-0.05	-0.07	-0.07	0.03	-0.02	0.02	-0.01	-0.02	-0.01	-0.01	0.01

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

Table A74.

Correlation between treatment effect estimate bias and percentage of clusters with pooled variance ratio <.5 or >2 with covariates weighted according to the strength of relation with the outcome, by ICC, cluster size, propensity score model, and matching method.

		Propensity score model and matching method															
Average	Cluster	RIS model			OP model			RI model			SL model			NoL2 model			No
ICC	size	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	P	2S	WC	match
.08	10	0.03	-0.01	0.14	0.07	0.03	0.03	0.02	0.01	-0.06	0.00	-0.02	0.09	0.00	0.04	0.01	-0.05
	25	0.01	0.06	0.08	0.03	0.11	0.03	0.03	0.00	-0.01	-0.01	0.05	0.09	0.06	-0.02	0.11	-0.04
	100	-0.01	0.11	-0.05	-0.01	0.00	-0.08	0.00	0.00	-0.07	0.02	0.10	-0.10	0.01	0.01	-0.06	0.02
	400	0.01	-0.03	-0.02	0.01	-0.01	-0.07	0.04	-0.01	0.04	-0.05	-0.06	0.05	0.05	-0.01	-0.04	0.03
.15	10	0.03	0.05	-0.01	-0.03	0.02	-0.05	-0.04	0.02	0.01	0.04	0.03	0.00	0.03	0.07	-0.01	-0.01
	25	0.00	0.02	0.12	0.00	0.00	0.02	0.02	0.06	-0.03	0.05	0.02	0.00	0.00	-0.07	-0.06	0.01
	100	0.03	0.05	-0.07	-0.05	0.08	0.01	0.03	-0.02	-0.01	0.12	0.00	-0.11	0.02	0.03	-0.09	0.01
	400	0.02	-0.03	-0.04	-0.07	-0.07	0.01	-0.08	-0.02	-0.03	0.06	0.08	-0.04	0.03	0.05	-0.06	0.07
.27	10	0.02	0.01	0.09	0.01	0.06	0.11	0.04	-0.02	0.05	0.01	0.01	0.00	-0.05	-0.05	0.03	0.05
	25	0.02	-0.01	-0.06	-0.01	0.00	-0.01	0.02	0.12	0.07	0.02	-0.06	0.04	-0.10	-0.07	0.07	-0.04
	100	-0.05	0.00	-0.04	0.05	-0.04	-0.03	0.01	-0.03	-0.06	-0.05	0.01	-0.05	0.04	-0.02	-0.04	-0.05
	400	-0.10	-0.02	0.03	-0.03	-0.02	-0.04	0.01	0.06	0.00	0.04	0.03	0.01	0.07	0.04	0.01	-0.05
.42	10	-0.02	0.00	0.08	-0.07	-0.01	0.06	0.06	0.00	0.08	-0.04	-0.01	0.04	-0.06	-0.04	0.04	0.03
	25	-0.03	0.02	0.02	0.05	0.10	0.04	0.01	0.03	0.00	0.09	0.07	0.01	0.08	0.10	0.01	-0.02
	100	-0.06	-0.06	0.04	0.00	-0.04	0.03	-0.02	-0.01	-0.02	0.01	-0.07	0.00	0.01	0.04	-0.04	0.02
	400	-0.01	-0.01	-0.03	0.05	-0.16	-0.10	-0.02	-0.03	0.05	0.01	0.01	0.01	-0.03	-0.01	0.01	0.01

Note. ICC=intracluster correlation of the unit-level covariates; RIS=random intercepts and slopes model; OP=over-parameterized model; RI=random intercepts model; SL=single-level model; NoL2=model without cluster-level covariates; P=pooled matching; 2S=two-stage matching; WC=within-cluster matching.

References

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine, 35,* 2074-2091.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*(4), 1770-1780.

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. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424.

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

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Statistics in Medicine,33*, 4306-4319.

Bell, B. A., Morgan, G. B., Kromrey, J. D., & Ferron, J. M. (2010). The impact of small cluster size on multilevel models: A Monte Carlo examination of two-level models with binary and continuous predictors. *JSM Proceedings, Survey Research Methods Section, 1*(1), 4057-4067.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H.H., de Boer, A., Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety, 20,* 1115-1129.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association, 91*(433), 109-122.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Second Edition.* New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research, 33,* 261-304.

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

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A, 128,* 134-155.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Craig, W., Herel-Fisch, Y., Fogel-Grinvald, H., Dostaler, S., Hetland, J., Simons-Morton, B., … & HBSC Bullying Writing Group. (2009). A cross-national profile of bullying and victimization among adolescents in 40 countries. *International Journal of Public Health, 54*(Suppl 2), 216-224.

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.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association, 94*(448), 1053-1062.

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

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*(9), 917-928.

Elstad, J. I., & Pedersen, A. W. (2012). The impact of relative poverty on Norwegian adolescents' subjective health: A causal analysis with propensity score matching. *International Journal of Environmental Research and Public, 9*(12), 4715-4731.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *Practice of Epidemiology, 173*(7), 761-767.

Gayat, E., Resche-Rigon, M., Mary, J., & Porcher, R. (2012). Propensity score applied to survival data analysis through proportional hazards models: A Monte Carlo study. *Pharmaceutical Statistics, 11*(3), 222-229.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology,41*(6), 872–884.

Gu, X., & Rosenbuam, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405-420.

Hammill, B. (2015). GMatch.SAS [SAS macro code]. Retrieved from http://people.duke.edu/~hammill/software/gmatch.sas.

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

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234-249.

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.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2017). MatchIt, Version 3.0.1.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis, 27*(3), 205-224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association, 101*(475), 901-910.

Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis, 29*(4), 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*(2), 407.

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

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*(1), 4-29.

Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review, 30*(3), 420-437.

Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis, 33*(4), 458-482.

Kiernan, K., Tao, J., & Gibbs, P. (2012). Tips and strategies for mixed modeling with SAS/STAT procedures. *SAS Global Forum, 332-2012.*

Kim, J., & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection processes vary across schools* (CSE technical report 708). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.

King, G., & Nielsen, R. (Forthcoming). Why propensity scores should not be used for matching. *Political Analysis.* Copy at http://j.mp/2ovYGsW.

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

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

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting and multilevel data. *Statistics in Medicine, 32*(19), 3373-3387.

Lian, Q., Su, Q., Li, R., Elgar, F. J., Liu, Z., & Zheng, D. (2018). The association between chronic bullying victimization with weight status and body self-image: a cross-national study in 39 countries. *PeerJ, 6*(4330).

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

Murnane, R. J., & Willett, J. B. (2011). *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* New York: Oxford University Press.

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

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological), 57(*1), 99-138.

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics, 39*(6), 612-636.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of repression coefficients when some regressors are not always observed. *Theory and Methods, 89*(427), 846-866.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational and Behavioral Statistics, 11*(3), 207-224.

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

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 68*(5), 688-701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6*(1), 34-58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*(371), 591-593.

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

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*(1), 249-264.

SAS [computer program]. (2014). Version 9.4. Cary, NC: SAS Institute Inc.

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

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

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125,* 303-353.

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), 1-21.

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,* S84-S90.

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

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.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*(3), 514-543.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., Hagedorn, M.C., Daly, P., & Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Codebook, Public Version* (NCES 2015-074). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (2014). Review Protocol for Beginning Reading Interventions (Version 3.0).

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (2017). What Works Clearinghouse: Standards Handbook (Version 4.0).

Waernbaum, I. (2010). Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference, 140,* 1948-1956.

Wagenaar, A. C., Maldonado-Molina, M., & Wagenaar, B. H. (2009). Effects of alcohol tax increases on alcohol-related disease mortality in Alaska: Time-series analyses from 1976 to 2004. American Journal of Public Health, 99(8), 1464-1470.

Winter, V. R., Combs, K. M., & Ward, M. (2018). An investigation of the association between foster care, body image, and BMI: A propensity score analysis. *Children and Youth Services Review, 84,* 82-85.

Wu, W., West, S. G., & Hughes, J. N. (2010). Effect of grade retention in first grade on psychosocial outcomes. *Journal of Educational Psychology, 102*(1), 135-152.