Title

Using a Two-Stage Propensity Score Matching Strategy and Multilevel Modeling to Estimate Treatment Effects in a Multisite Observational Study

Author

Rickles, Jordan

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Using a Two-Stage Propensity Score Matching Strategy and Multilevel Modeling to Estimate Treatment Effects in a Multisite Observational Study

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Education

by

Jordan Harry Rickles

2012

ABSTRACT OF THE DISSERTATION

Using a Two-Stage Propensity Score Matching Strategy and
Multilevel Modeling to Estimate Treatment Effects
in a Multisite Observational Study

by

Jordan Harry Rickles
Doctor of Philosophy in Education
University of California, Los Angeles, 2012
Professor Michael Seltzer, Chair

In this study I present, demonstrate, and test a method that extends the Stuart and Rubin (2008) multiple control group matching strategy to a multisite setting. Three primary phases define the proposed method: (1) a design phase, in which one uses a two-stage matching strategy to construct treatment and control groups that are well balanced along both unit- and site-level key pretreatment covariates; (2) an adjustment phase, in which the observed outcomes for non-local control group matches are adjusted to account for differences in the local and non-local matched control units; and (3) an analysis phase, in which one estimates average causal effects for the treated units and investigates heterogeneity in causal effects through multilevel modeling. The main novelty of the proposed method occurs in the design phase, where propensity score matching is executed in two stages. In the first stage, treatment units are matched to control units

within the same site. In the second stage, treatment units without an acceptable within-site match are matched to control units in another site (between-site match). The two-stage matching method provides researchers with an alternative to strict within-site matching or matching that ignores the nested data structure (pooled matching). I employ an empirical illustration and a set of simulation studies to test the utility and feasibility of the proposed two-stage matching method. The results document the two-stage matching method's conceptual appeal, but indicate that effect estimation under the two-stage matching method does not, in general, outperform more traditional matching-based or regression-based methods. Alternative specifications within the proposed method can improve performance of two-stage matching. In addition to extending the work of Stuart and Rubin, this study complements the small set of studies that have examined propensity score matching in multisite settings and provides guidance for researchers looking to estimate treatment effects from a multisite observational study. The dissertation concludes with directions for future research and considerations for researchers conducting multisite observational studies.

The dissertation of Jordan Harry Rickles is approved.

Thomas Belin

Li Cai

Noreen Webb

Michael Seltzer, Committee Chair

University of California, Los Angeles

2012

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation would not have materialized without the teaching, guidance, collaboration and support of countless individuals. My gratitude starts with my dissertation committee members. My advisor, Michael Seltzer, has been a positive motivator throughout my doctoral studies. I am grateful to have him as a teacher, mentor and friend. I truly appreciate the time he devoted to help guide me through the dissertation process. I am also thankful to have had Noreen Webb's supportive presence throughout my time in the Social Research Methodology program, from the first quarter through to the completion of this dissertation. Throughout my post-secondary education—eleven years long—I took more courses (five) taught by Li Cai than any other professor. I could not have executed a significant portion of my analysis without the skills he imparted. Similarly, I must thank Tom Belin for exposing me to causal inference as a field of study, the foundations of which permeate this dissertation. I also want to acknowledge the academic lineage of my committee members: Anthony Bryk (Michael Seltzer's advisor), Lee Cronbach (Noreen Webb's advisor), Donald Rubin (Tom Belin's advisor), and David Thissen (Li Cai's advisor). I am humbled and honored to even have a tangential connection with those great minds, and I hope it is clear in this dissertation that their work has greatly shaped my thinking either directly or indirectly.

I would also like to extend my thanks to other faculty who helped guide my learning and provided support during this process: Mike Rose for his inspiration to think not just about the data but the people represented by the data; Meredith Phillips for first exposing me to the Campbell and Stanley framework and being a willing mentor well beyond my time in the Public Affairs building; and Elizabeth Stuart for taking her time to listen to my ideas and provide guidance whenever our paths crossed at conferences. My graduate school experience and

dissertation also benefited from numerous conversations with my Social Research Methodology peers. I would like to specifically thank the following colleagues for their intellectual support and stimulation: Mark Hansen, Scott Monroe, Jon Schweig, Larry Thomas, and Ji Seung Yang. Ji Seung deserves particular acknowledgment for helping me debug my matrix algebra code.

I owe my greatest thanks to my family for their patients and support. To my parents and in-laws, thank you for making our lives easier and providing help whenever it was needed. To Ian and Maya, who provided constant motivation to finish and reinforced the importance of multitasking. And most importantly, to my better half, Yukari, for always being a positive, encouraging voice. Without her support this would not have been possible, or worthwhile.

Lastly, I want to acknowledge the funding support I received during the past four years. Part of this research was made possible by a pre-doctoral advanced quantitative methodology training grant (#R305B080016) awarded to UCLA by the Institute of Education Sciences of the US Department of Education. The views expressed in this paper are mine alone and do not reflect the views/policies of the funding agencies or grantees.

Vita

EDUCATION

1998	B.S., Industrial and Labor Relations School of Industrial and Labor Relations Cornell University Ithaca, NY
2000	M.P.P., Public Policy Luskin School of Public Affairs University of California, Los Angeles Los Angeles, CA

WORK EXPERIENCE

2000-2005	Staff Research Associate Ralph and Goldy Lewis Center for Regional Policy Studies University of California, Los Angeles Los Angeles, CA
2002-2003	Lecturer Department of Political Science California State University, Los Angeles Los Angeles, CA
2005-2008	Educational Research Analyst Los Angeles Unified School District Los Angeles, CA
2008-2011	Graduate Student Researcher National Center for Research on Evaluation, Standards, and Student Testing University of California, Los Angeles Los Angeles, CA
2010-2012	Special Reader (Teaching Assistant) Department of Education University of California, Los Angeles Los Angeles, CA

SELECT PUBLICATIONS

2004	Ong, P. M. & Rickles, J. The Continued Nexus between School and Residential Segregation. In *Symposium, Rekindling the Spirit of Brown v. Board of Education, California Law Review*.
2005	Rickles, J. & Ong, P. M. The Integrating (and Segregating) Effect of Charter, Magnet, and Traditional Elementary Schools: The Case of Five California Metropolitan Areas, *Journal of California Politics and Policy*: June.
2011	Rickles, J. Using Interviews to Understand the Assignment Mechanism in a Non-Experimental Study: The Case of 8th Grade Algebra. *Evaluation Review*, 35 (5).
2012	Seltzer, M. & Rickles, J. Multilevel Analysis. In J. Arthur, M. Waring, R. Coe and L. Hedges (Eds.), *Research Methods and Methodologies in Education*. London: Sage Publications.
Forthcoming	Rickles, J. Examining Heterogeneity in the Effect of Taking Eighth-Grade Algebra on High School Mathematics Achievement. *Journal of Educational Research*.

AWARDS AND HONORS

2008-2012	Institute of Education Sciences Predoctoral Fellowship Advanced Quantitative Methods, Department of Education University of California, Los Angeles Los Angeles, CA
2010	Leigh Burstein Award for Excellence in Social Research Methodology Graduate School of Education and Information Studies University of California, Los Angeles Los Angeles, CA

Chapter 1

Introduction

Does participation in extracurricular activities help keep students in school (McNeal, 1995)? Does school suspension affect future academic success (Fabelo et al., 2011)? Does grade retention improve long-term student outcomes (Hong & Raudenbush, 2005)? Does participation in an advanced placement high school course improve college outcomes (Hardgrove, Godin, & Dodd, 2008)? Does taking algebra in middle school affect long-term mathematics attainment (Smith, 1996)? Besides being of educational importance, questions like these share a common set of methodological characteristics that complicate educational research.

First, questions like those above invoke an interest in the causal effect of an educational policy, program, or practice on student outcomes. As such, one could make unbiased causal inferences regarding these questions with a randomized experimental study design. Despite increased use of experimental designs in educational research (Shadish & Cook, 2009), however, a controlled experiment is often not feasible for ethical, practical, or political reasons. While simple direct comparisons of treated and untreated groups will likely result in biased effect estimates, one can use non-experimental, or observational, methods to address these causal questions.

Second, these questions ask about policies, programs or practices implemented within multiple sites, e.g., schools, where the policies/programs/practices target a select population and implementation can vary across sites. Since assignment to the treatments in question can be highly selective and this selectivity can differ across sites, standard methods for selection bias adjustment may be ineffective and/or inefficient. For example, regression-based approaches may be highly dependent on parametric assumptions and extrapolation, and matching-based

approaches may be limited by a lack of covariate overlap and common support. Lastly, the effects from any one treatment could differ across students and schools, and thus any findings based on overall average effects might mask more meaningful information stemming from effect heterogeneity (Cronbach, 1976).

In this study I present, demonstrate, and test a method designed to facilitate estimation of causal effects and effect heterogeneity for research questions, like those above, that manifest in a multisite observational setting. By multisite, I mean situations where treatment conditions (e.g., taking an advanced mathematics course or standard mathematics course) are assigned to units (e.g., students) within different sites (e.g., schools). One can consider each site as a separate mini-study within the larger multisite study (Seltzer, 2004). This is in contrast to single-site studies where units and treatment assignment are not nested within sites, and in contrast to cluster design studies where treatments are assigned at the site or cluster level and not to individual units within sites (Raudenbush, 1997). By observational, I mean a situation where the study design and data collection are conducted after treatment conditions are assigned to units and the mechanism by which treatment conditions are assigned to units is not random nor completely known by the researcher (Rosenbaum, 2002). Some make a slight distinction between observational and quasi-experimental studies, where in a quasi-experimental study the researcher has some control over treatment assignment and/or data collection but random assignment is not possible (Shadish, Cook, & Campbell, 2002). I refer to both observational and quasi-experimental studies as non-experimental because they lack randomization. I use the terms observational and quasi-experimental interchangeably throughout this paper. Therefore, this study focused on multisite observational settings where researchers have access to pre-existing

data for multiple sites and seek to make causal inferences about a treatment that units within sites either select or are assigned in a way that is at least partially unknown to the researcher.

The proposed method is a multisite extension of a single-site method described in Stuart and Rubin (2008), in which the analysis is separated into three phases: a design phase for preprocessing the data, an adjustment phase for adjusting for possible between-site bias, and an analysis phase for estimating average treatment effects and effect heterogeneity. The main novelty lies in the design phase, where a two-stage matching approach is used to select appropriate control units for the treatment units. The method is rooted in the potential outcomes framework for causal inference (Rubin, 1974, 2005) and utilizes multilevel, or hierarchical, modeling (Raudenbush & Bryk, 2002) and propensity score matching (Rosenbaum & Rubin, 1983, 1984, 1985) to adjust for selective pre-treatment differences between treated and untreated units.

Despite widespread use of the potential outcomes framework and multilevel modeling for educational research over the past decade, understanding the nexus of these two methodological advances is still in its infancy. Literature on causal inference and the potential outcomes framework often ignores, or avoids, discussion of a multilevel context, and literature on multilevel modeling often ignores, or avoids, discussion of causal inference outside of a randomized control trial. It is important to gain a better understanding of the complications and considerations researchers must address when trying to draw causal inferences from non-experimental multilevel data. For example, the appropriate application of propensity score techniques in a multilevel setting has been the subject of considerable research interest over the last few years (Arpino & Mealli, 2011; Kelcey, 2011b; Kim & Seltzer, 2007; Steiner, 2011; Su & Cortina, 2009; Thoemmes, 2009; Thoemmes & West, 2011) because one must consider not only

unit-level confounding but also site-level confounding and cross-level interactions. Within-site selectivity of treatment assignment further complicates propensity score matching techniques because one must make trade-offs between prioritizing matching on unit-level or site-level criteria. Additionally, concerns about parametric assumptions and extrapolation with regression-based adjustments for effect estimation can be compounded with multilevel models, where these assumptions need to hold within each level of analysis.

This study helps shine light on the complications and considerations that arise when trying to draw causal inferences from non-experimental multilevel data, as well as demonstrate a method that builds on both multilevel modeling and propensity score matching to address some of the complications. Paired with past research, the study's findings provide some guidance for researchers looking to estimate treatment effects from a multisite observational study. Most importantly, researchers should take efforts to understand what factors influence both treatment assignment and outcomes of interest, giving particular attention to the importance of site-level factors relative to unit-level factors. When site-level factors play an important role in treatment assignment, within-site matching is preferred. The two-stage matching method provides an alternative option when within-site matching is limited.

In the following chapter I discuss the conceptual foundation and literature from which the proposed method builds upon. In chapter three, I describe the proposed method in detail, outline the research questions that guided the study, and describe the methods I used to address those questions. Namely, I used an empirical data set to illustrate the execution and utility of the proposed two-stage matching method and a set of simulation studies to examine how the method performs under different specifications. Results from the empirical illustration are presented in

Chapter 4, while the simulation study results are presented in Chapter 5. I conclude the study with a discussion of the findings and implications for researchers.

Chapter 2

Conceptual Framework and Literature Review

The proposed method for estimating causal effects in a multisite observational setting utilizes some well-established statistical techniques (e.g., hierarchical modeling and propensity score matching) and builds off of research design concepts developed in recent years to address complications that arise when the research setting strays from simplified textbook examples. First and foremost, however, the proposed method is rooted in the potential outcomes framework for causal inference popularized by Rubin (1974). As such, the first part of this chapter describes the key features and implications of the potential outcomes framework for estimating causal effects. Within this framework, I briefly review the main research design and statistical approaches used for effect estimation when treatments are contained within a single site. I then discuss the literature that arose to address complications in the potential outcomes framework when one wishes to make causal claims in a multisite setting, with particularly emphasis on settings within education. I conclude the chapter with a summary of the key concepts, methods, and complications that motivated the proposed method, and how the proposed study contributes to both the fields of causal inference and educational research.

2.1. The potential outcomes framework

Under the potential outcomes framework, causal effects are defined at the individual unit of analysis (e.g., a student) as the difference between the unit's outcome under a treatment condition and what that unit's outcome would have been in the absence of the treatment. Holland (1986) traces this general notion of a causal effect back to, most notably, John Stuart Mill and

the statistician Jerzy Neyman. This definition of a causal effect can also be traced back to writings in economics (Roy, 1951). The framework was formalized and popularized, however, by Rubin in a series of articles (Rubin, 1974, 1976, 1978, 1980), and as a result, is often referred to as the Rubin Causal Model. While other frameworks for causal inference exist (Dawid, 2000; Greenland & Brumback, 2002), the potential outcomes framework is embedded in the research methodology literature for a wide variety of disciplines, including: education (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007), economics (Angrist & Pischke, 2008; Heckman, 1989), sociology (Morgan & Winship, 2007), political science (Ho, Imai, King, & Stuart, 2007), public health (Little & Rubin, 2000), and more general literature on research design (Shadish et al., 2002).

A key feature of the potential outcomes framework is that the causal effect of a given treatment for an individual unit, *i*, can be defined mathematically based on the unit's potential outcomes under the different treatment conditions (for simplicity, assume two treatment conditions):

$$\delta_i = Y(1)_i - Y(0)_i \tag{2.1}$$

This definition emphasizes the fact that one can only calculate the true causal effect for a given treatment on a given outcome for a given unit, δ_i, if one can observe the outcome for that unit when exposed to the treatment condition, $Y(1)_i$, and the outcome for that unit when exposed to the alternative condition (i.e., the control condition), $Y(0)_i$. Therefore, causal effects at the unit level are defined by a unit's potential outcomes: $Y(1)_i$ and $Y(0)_i$.

2.1.1. The stable-unit-treatment-value assumption

The units, treatment conditions, and potential outcomes comprise the causal effect quantities that one wishes to estimate. Rubin (2005) refers to these quantities as causal estimands and, more generally, the definition of these quantities as the "science" one wishes to learn about. To conceptualize and define potential outcomes and a given estimand, it is often necessary to invoke the "stable-unit-treatment-value" assumption (SUTVA) (Rubin, 1978, 1980). For SUTVA to hold, two criteria must be met. First, the potential outcomes for a given unit must not depend on the treatment received by another unit. In other words, there cannot be any interference between units, so the potential outcomes for unit A will be the same regardless of whether unit A receives the same treatment as unit B or not. Second, the treatment condition for which a unit is assigned must be invariant. In other words, if unit A and unit B are assigned to treatment 1, they will experience the same treatment. When it comes to studies in education, the plausibility of SUTVA is tenuous—an issue I address later in this chapter. Without SUTVA, the number of potential outcomes for a given unit can increase exponentially, and thus make it difficult to even conceptualize, let alone estimate, a causal effect of interest.

2.1.2. The assignment mechanism

Assuming one can posit a set of potential outcomes and an estimand of interest, one can only observe a single potential outcome for the same unit under the same exact conditions (e.g., at the same point in time). This reality is often referred to as the "fundamental problem of causal inference" (Holland, 1986). For example, for a student assigned to a treatment group (D_i=1), we will observe $Y(1)_i$, and $Y(0)_i$ only exists under an unobserved counterfactual condition.

Conversely, for a student assigned to a control group (D_i=0), we will observe $Y(0)_i$, and $Y(1)_i$ only exists under an unobserved counterfactual condition.

While unobserved potential outcomes preclude us from estimating causal effects for an individual unit, we can calculate average causal effects across units in a given sample based on the expectations (e.g., means) of the observed outcome for each treatment group:

$$\mathrm{E}[Y(1) \mid D=1] - \mathrm{E}[Y(0) \mid D=0] \tag{2.2}$$

Estimates of average causal effects based on Equation 2.2 result in what Holland (1986) called the "prima facie causal effect" (p. 949) and what Morgan and Harding (2006) called the "naïve estimator of the average causal effect" (p. 10). Whether such estimates are unbiased, however, depends on the process by which the units of analysis select, or are assigned, to the treatment condition(s) of interest. Rubin (1991, 2004a, 2005) emphasizes this point and argues that statistical inference for causal effects "requires the specification of a posited assignment mechanism describing the process by which treatments were assigned to units" (1991, p. 403).

It is this assignment mechanism that determines whether we observe either $Y(0)_i$ or $Y(1)_i$. For Equation 2.2 to hold as an unbiased estimate of an average treatment effect (ATE), assignment to treatment conditions must be independent of the potential outcomes: $[Y(1) , Y(0)] \perp D$. In statistics, this is commonly referred to as the selection independence assumption (Holland, 1986). Under selection independence, $\mathrm{E}[Y(1) \mid D=1] = \mathrm{E}[Y(1) \mid D=0] = \mathrm{E}[Y(1)]$ and $\mathrm{E}[Y(0) \mid D=0] = \mathrm{E}[Y(0) \mid D=1] = \mathrm{E}[Y(0)]$, so Equation 2.2 will equal $\mathrm{E}[Y(1) - Y(0)]$ and will be an unbiased estimate of the ATE. The intent and appeal of random assignment is that the randomization to treatment and control conditions is an assignment mechanism designed to produce selection independence.

If selection independence does not hold, however, Equation 2.2 will produce biased estimates of the ATE. Winship and Morgan (1999) show how the degree to which the ATE defined by Equation 2.2 represents the true average treatment effect depends on "the way in which individuals are assigned (or assign themselves) to the treatment and control groups" (p. 665). Shadish et al. (2002) discuss this bias as different threats to internal validity, most notably selection bias, where pre-existing differences in the treatment and control groups mean one group would do better than the other regardless of treatment receipt so $E[Y(0) \mid D=0] \neq E[Y(0) \mid D=1]$. For example, student assignment to an above-grade-level math course might be based on a math test taken at the end of the previous year. If this is the case, then a causal estimate of the above-grade-level math course effect based on Equation 2.2 will ignore the fact that students with higher pretest scores are more likely to be exposed to the treatment and more likely to have higher potential outcomes regardless of treatment assignment.

The difference between a situation where the assignment mechanism is independent of the potential outcomes and a situation where the assignment mechanism is at least partially based on a covariate, such as a pretest, can be depicted through graphical representation (Pearl, 2009). Figure 2.1 shows the relationship between D, Y, X (an observed covariate), and U (an unobserved covariate) under three differ assignment mechanisms: random assignment (panel A), assignment based on an observed covariate (panel B), and assignment based on an observed and unobserved covariate (panel C). Arrows from one variable to another indicate a causal path. If randomization is employed to determine who is assigned to treatment and control conditions, then $(X,U) \perp D$ and treatment and control groups should have similar covariate distributions (e.g., $E[X \mid D=1] = E[X \mid D=0]$). Additionally, treatment and control groups should have similar expected potential outcomes under the control condition (i.e., $E[Y(0) \mid D=0] = E[Y(0) \mid D=1]$), which implies that the

difference in the observed outcomes, E[$Y(1) \mid D=1$] - E[$Y(0) \mid D=0$], is due to the treatment effect and not preexisting group differences.

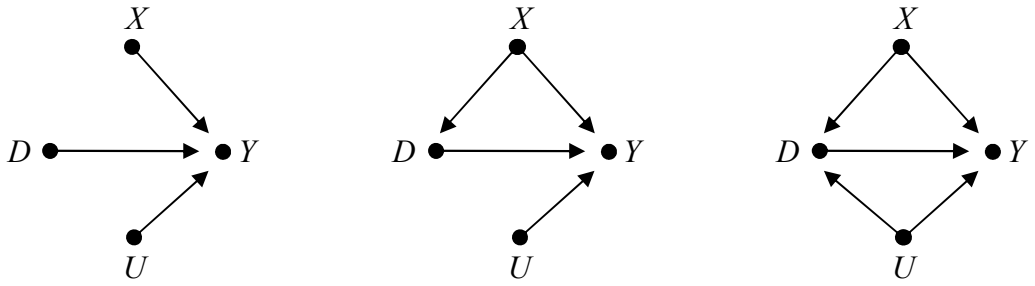

Figure 2.1. Three different types of assignment mechanisms.

If, however, a covariate partially determines treatment group assignment (as in panels B and C in Figure 2.1), then treatment and control groups will not, in general, have the same covariate distributions (e.g., E[$X \mid D=1$] ≠ E[$X \mid D=0$]). Under this type of assignment mechanism, treatment and control groups will not have similar expected potential outcomes under the control condition (i.e., E[$Y(0) \mid D=0$] ≠ E[$Y(0) \mid D=1$]), and differences in the observed outcomes could be due to either the treatment, differences in the covariate(s), or both. An analysis that ignores the covariate(s) will falsely attribute the entire difference to the treatment effect. If, however, treatment assignment is conditionally independent of the potential outcomes given the observed covariates (i.e., [$Y(1)$, $Y(0)$] $\perp$ $D \mid X$), one can estimate an average treatment effect from the following conditional expectations (Heckman & Hotz, 1989; Rosenbaum & Rubin, 1983):

$$E[Y(1) \mid D=1, X] - E[Y(0) \mid D=0, X] \quad (2.3)$$

The assignment mechanism depicted in panel B of Figure 2.1 is a case where conditional independence holds. Heckman and Hotz (1989) refer to this type of assignment mechanism as selection on the observables because treatment assignment only depends on factors the researcher measured and can condition on. Rosenbaum and Rubin (1983) refer to this situation as ignorable treatment assignment because the counterfactual potential outcomes are ignorable given the observed pretreatment differences between treatment and control groups. In other words, X is a vector of covariates that includes all the pretreatment factors that partially determine both treatment assignment and the potential outcomes. For treatment and control units with the same discrete value of X, treatment assignment is independent of all other pretreatment factors (i.e., $U \perp D$) and differences in observed outcomes can be attributed to the treatment. For example, if students who passed a pretest were more likely to take an above-grade-level math course than students who failed—and no other factors related to academic performance influenced course assignment—then outcome comparisons within pretest pass/fail categories could produce unbiased treatment effect estimates. If, however, the assignment mechanism is based on both observable and unobservable covariates (Figure 2.1, panel C), then treatment effect estimates based on Equation 2.3 will, like estimates based on Equation 2.2, fail to produce unbiased estimates.

Note that for causal inference, Rosenbaum and Rubin (1983) extend the notion of ignorable treatment assignment to a strongly ignorable treatment assignment assumption, where in addition to selection on the observables, treatment assignment for any given individual unit in the analysis cannot be deterministic. In other words: $0 < P(D_i = 1) < 1$. Also note that the graphical diagrams in Figure 2.1 represent a simplified depiction of the data, and does not capture important nuances such as strongly ignorable treatment assignment. Furthermore, in

practice, multiple observed and unobserved covariates may exist, along with interactions among the covariates, which are difficult to convey through the graphical representations.

The assignment mechanism will often determine the type of analyses one can conduct and the plausibility of any causal inferences drawn from those analyses. If selection independence holds, as in a randomized controlled trial, treatment effect estimation is relatively straightforward based on Equation 2.2. If selection independence does not hold, as in most non-experimental research conditions, more complex research designs and statistically modeling techniques might be required to condition on the observable covariates and generate treatment effect estimates following Equation 2.3. Estimates based on Equation 2.3, however, require an assumption of strongly ignorable treatment assignment, and the plausibility of this assumption depends on whether the observables controlled for in the analysis adequately capture the assignment mechanism. A more conscious effort to address the assignment mechanism can go a long way to addressing the confidence one places in causal effect estimates in a non-experimental study. Cook, Shadish, and Wong (2008) reviewed within-study comparisons to identify the conditions under which non-experimental studies can produce estimates comparable to experiments. They found that "[k]nowledge of the selection process can significantly reduce selection bias provided the selection process is valid and reliably measured" (p. 740). Similarly, Shadish, Clark, and Steiner (2008) concluded "that attention to careful measurement of the selection process can be crucial to the success of subsequent analyses" (p. 1341).

2.1.3. Causal inference as a missing data problem

Formulating causal effects with the potential outcomes framework makes it clear that efforts to estimate causal effects are efforts to address missing data. The connection between

missing data, potential outcomes, and causal inference was first made by Rubin (1978) when formulating causal inference from a Bayesian perspective. He noted that "problems of inference for causal effects of treatments on individual experimental units or collections of experimental units are equivalent to problems of inference about values of missing data" (p. 39).

Just as the validity of causal inferences hinge on the assignment mechanism, valid inferences with missing data hinge on the missingness, or response, mechanism that indicates whether a given variable is observed or not observed for a given unit (Rubin, 1976; Schafer & Graham, 2002). Based on the response mechanism, missing data can take one of three forms: missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Data are MCAR when the response mechanism does not depend on the missing or observed data, which is analogous to situations where selection independence holds and treatment assignment does not depend on the potential outcomes. Under such conditions, methods such as imputing missing data with unconditional means can result in unbiased estimates of population means, just as Equation 2.2 holds under selection independence. For example, one could impute the counterfactuals for students assigned to treatment based on unconditional mean substitution by setting $Y(0)_i = \mathrm{E}[Y(0) \mid D=0]$, although bringing in additional information could improve precision of the imputation. Data are MAR when the response mechanism depends on the observed data but not the missing data, which is analogous to situations where strongly ignorable treatment assignment holds. Under MAR, methods such as imputing missing data with means conditional on the observed data can result in unbiased estimates of population means, just as Equation 2.3 holds under strongly ignorable treatment assignment. For example, one could impute the counterfactuals for students assigned to treatment based on conditional mean substitution by setting $Y(0)_i = \mathrm{E}[Y(0) \mid D=0, X_i]$. Data are NMAR

when the response mechanism depends on the missing data, which is analogous to situations where treatment assignment depends on both observable and unobservable factors. Under NMAR, imputing missing data will not result in unbiased estimates of population means. An overview of methods for handling missing data can be found in, for example, Allison (2001), Little and Rubin (2002) and Schafer and Graham (2002).

In a non-experimental setting, counterfactual potential outcomes are not likely to be MCAR but may be MAR. Some approaches to treatment effect estimation address the "fundamental problem of causal inference" through methods that impute the counterfactual outcomes. For example, Schafer and Kang (2008) discuss using regression estimated values for the counterfactual outcomes, an approach employed by Reinisch et al. (1995) to study the effect of in utero exposure to a prenatal drug on intelligence as an adult. Additionally, a few studies (Hirano, Imbens, Rubin, & Zhou, 2000; Imbens & Rubin, 1997; Taylor & Zhou, 2009) take a more Bayesian approach and multiply impute counterfactuals. Conceptually, multiple imputation is appealing because it aligns with Rubin's (1978) formulation of causal inference from a Bayesian perspective. From this perspective, one can draw individual counterfactual potential outcome values from a posterior predictive distribution. In fact, Rubin (2004b) argued that "all problems of causal inference should be viewed as problems of missing data: the potential outcomes under the not-received treatment are the missing data. A straightforward and valid way to think about missing data is to think about how to multiply impute them" (p. 167).

2.2. Common methods for causal effect estimation

The previous section provided a conceptual framework for defining causal effects and understanding the strengths and weaknesses of different methods for estimating causal effects.

These methods typically apply a combination of research design techniques that facilitate the comparison of units exposed to at least two different treatment conditions (e.g., treatment and control) and statistical modeling techniques to estimate treatment effects. Both research design and statistical methods for causal inference have been well documented by, for example, Shadish et al. (2002), Morgan and Winship (2007), Rosenbaum (2009), and Schafer and Kang (2008). In this section, I review the main methods as they pertain to single-site observational studies, and highlight particular empirical examples in education that are relevant to questions about the effects of curriculum differentiation. I begin the section, however, by briefly discussing the importance of research design and appeal of randomized experiments.

2.2.1. The importance of design

While not clearly distinct in practice, methods for causal inference can be divided into two broad stages: research design and analysis. To infer that a given treatment, D, has a causal effect on a given outcome (Y), one must determine that D and Y are correlated and that the correlation represents a causal relationship from D to Y. Uncovering a relationship between D and Y is primarily undertaken in the analysis stage and can be assessed by what Shadish et al. (2002) refer to as statistical conclusion validity. Inferring that the observed relationship is causal depends on internal validity. The degree to which a causal inference can withstand threats to internal validity will depend, to a large extent, on the research design.

The importance of design is prevalent in the literature that builds from the work conducted in the 1950s and 1960s by both Donald Campbell and William Cochran. For example, in discussing the "Campbell Causal Model" (CCM), Shadish (2010) wrote that one key feature is "the use of validity types and threats to analyze and prevent likely inferential problems in the

design of cause-probing studies, both retrospectively and especially prospectively" (p. 4) and went on to say that "CCM emphasizes the primacy of design over analysis" and "the first line of attack toward good causal inference is to design studies that reduce 'the number of plausible rival hypotheses available to account for the data.'" (p. 5). Similarly, in discussing Cochran's contributions to observational studies, Rubin (2006) identified design as a major theme. For design, Cochran emphasized "the need to measure, as well as possible, important variables … the need for a control group … the desire to avoid control groups with large initial biases" and the use of "matched sampling or blocking to reduce initial bias" (Rubin, 2006, p. 10). The utility of designing studies that use matching or blocking was also discussed in the 1950s by the sociologist Samuel Stouffer, who "advocated matching designs to select subsets of seemingly equivalent individuals from those who were and were not exposed to the treatment" (Morgan & Winship, 2007, p. 8).

Both the potential outcomes framework, as discussed above, and the concept of internal validity threats, as discussed by Shadish et al. (2002), make clear the appeal and utility of randomized experimental designs. Since one can only observe one potential outcome after treatment assignment, causal inference depends on constructing a counterfactual value through comparison to another group of units, and the validity of any comparison will depend on the assignment mechanism. A randomized experiment, if implemented correctly with full compliance and no attrition, should produce an assignment mechanism with selection independence, where all observed and unobserved covariates will be unrelated to treatment assignment (i.e., Figure 2.1a). By creating treatment and control groups that are equivalent just prior to treatment exposure, threats to internal validity (e.g., selection bias) can be ruled out. Under such conditions, estimating a causal effect in the analysis stage can be as simple as

comparing two group means. As a result, a design based on random assignment is the preferred method for estimating causal effects.

When a randomized design is not possible, it is useful to follow the advice of Dorn and ask, "How would the study be conducted if it were possible to do it by controlled experimentation?" (Cochran, 1965, p. 236). In other words, the nonrandomized design should try to mimic as closely as possible the hypothetical randomized design one would like to conduct. The main purpose of a randomized design is to create a control group (or groups) as similar as possible to the treatment group (Shadish & Cook, 2009), so a key goal for a nonrandomized design is to construct a control group that mirrors the treatment group in all respects except for treatment exposure. As Cochran (1965) wrote, the ideal control group is one that "while lacking the suspected causal factor, should have the same distribution as the chosen study group with regard to all major disturbing variables" (p. 248). Similarly, Shadish and Cook (2009), paraphrasing advice given years earlier by Campbell, recommend that "the desirable control in a [nonrandomized study] is a focal local control group: in the same locale as the treatment group and focused on persons with the same kinds of characteristics as those in the treatment group, most particularly the characteristics that are most highly correlated with selection into conditions and with the outcome under investigation" (p. 619).

The utility of a control group depends on whether it allows the researcher to rule out rival hypotheses for the relationship between treatment and outcome. These rival hypotheses are most likely to stem from confounding, or what Cochran referred to as disturbing variables. For example, in studying the effect of taking an above-grade-level course on subsequent mathematics achievement, factors such as prior mathematics achievement and interest in mathematics are likely to confound any observed relationship between course taking and achievement. In a

randomized design, treatment assignment will be independent of both observed and unobserved potential confounding variables, but in a nonrandomized design one can only work with observed variables. Thus, a good design will identify and measure the confounding variables and construct a control group that, as with a randomized design, results in no association between the confounding variables and treatment receipt (i.e., results in strongly ignorable treatment assignment). Cochran (1965) recommended constructing a list of disturbing variables arranged into three classes:

(1) major variables for which some kind of matching or adjustment is considered essential;
(2) variables for which we would like to match or adjust but can just verify that their effects produce little or no bias; and
(3) variables whose effects are minor and can be disregarded.

Applying these three classes to the above-grade-level example, we would want a focal local control group that is focal in the sense that the control group has the same distribution of prior mathematics achievement and interest in mathematics (disturbing variables in Cochran's class one) as the treatment group and local in the sense that the students attend the same school (or district). The emphasis on a local control group helps rule-out confounding factors having to do with group membership, e.g. school-level learning environment or peer effects, that may fall into Cochran's class one or two list of disturbing variables.

Another distinct feature of a randomized experimental design is that the design and data collection precede, and dictate, the analysis. Rubin emphasizes this point when discussing the importance of design for observational studies: "[L]ike good experiments, good observational studies are designed, not simply 'found.' When designing an experiment, we do not have any outcome data, but we plan the collection, organization, and analysis of the data to improve our

chances of obtaining valid, reliable, and precise causal answers. The same exercise should be done in an observational study" (Rubin, 2004a, p. 356). By formally setting up the control group prior to examining the outcomes, one can avoid fishing for a desired answer through repeatedly re-specifying the design and analysis. In other words, "lack of availability of outcome data when designing experiments is a tremendous stimulus for 'honesty' in experiments and can be in well-designed observational studies as well" (Rubin, 2001, p. 366).

In the following two sub-sections I discuss the two main methods for causal effect estimation in an observational study: regression-based approaches and matching-based approaches. These two approaches differ in the degree to which the design stage is prioritized over the analysis stage, and the differences can have implications for the validity of causal effect estimates. When examining these two approaches, it is important to keep in mind that "a fundamental element of good quasi-experimental design is that a focal local control makes the job of estimating causal effects much easier from the start" (Shadish & Cook, 2009, p. 619).

2.2.2. Regression-based approaches to causal effect estimation

Regression-based, or analysis of covariance (ANCOVA), methods were originally applied to causal inference to increase the precision of causal effect estimates in randomized experiments. In a nonrandomized study, regression-based methods are employed to remove bias in the causal effect estimate by adjusting for pre-existing differences between the treatment and control groups. Since the 1960s, regression-based approaches have been the most widely used method for estimating causal effects in nonrandomized studies (Morgan & Winship, 2007), and have been extensively applied to estimate treatment effects related to curriculum differentiation. For example, the effects of ability grouping (Burks, 1994; Hallinan & Kubitschek, 1999; Hoffer,

1992) and timing of algebra course taking (Gamoran & Hannigan, 2000; Ma, 2005; Smith, 1996) have been examined with regression-based adjustments for pre-existing group differences.

The logic and potential pitfalls of the regression-based approach can be demonstrated by returning to the above-grade-level example. Under a situation where assignment to an above-grade-level mathematics course (*D*) is partially determined by an observed prior mathematics achievement test score (*X*), one can estimate the treatment effect on an outcome of interest (*Y*) with the following ordinary least squares (OLS) regression model:

$$Y_i = \beta_0 + \beta_1(X_i - \bar{X}_.) + \delta D_i + e_i,$$

where $(X_i - \bar{X}_.)$ is the grand-mean centered prior mathematics achievement test score for student *i*, and individual deviations from the expected outcome value given *D* and *X* are capture by e_i, which is assumed to be normally distributed with mean zero and standard deviation σ^2. The regression coefficients β_0, β_1, and δ are parameters to be estimated, with $\hat{\delta}$ representing the estimated average treatment effect as defined in Equation 2.3. By rearranging the terms in the estimated regression model and substituting group averages for individual observed values, we can see how $\hat{\delta}$ differs from the naïve average treatment effect defined in Equation 2.2 (Cochran & Rubin, 1973; Gelman & Hill, 2006). For a given sample, the observed average outcome for the treatment group ($\bar{Y}^t$) is defined as $\bar{Y}^t = \hat{\beta}_0 + \hat{\beta}_1(\bar{X}^t - \bar{X}) + \hat{\delta}$, where $\bar{X}^t$ is the average prior mathematics score for the treatment group. Similarly, the observed average outcome for the control group ($\bar{Y}^c$) is defined as $\bar{Y}^c = \hat{\beta}_0 + \hat{\beta}_1(\bar{X}^c - \bar{X})$, where $\bar{X}^c$ is the average prior mathematics score for the control group. Therefore,

$$\bar{Y}^t - \bar{Y}^c = \hat{\beta}_1(\bar{X}^t - \bar{X}^c) + \hat{\delta}$$

and

$$\hat{\delta} = \bar{Y}^t - \bar{Y}^c - \hat{\beta}_1(\bar{X}^t - \bar{X}^c). \quad (2.4)$$

Equation 2.4 shows that regression-based adjustment for the prior mathematics test score, depends on two ingredients: the average prior test score difference between treatment and control groups $(\bar{X}^t - \bar{X}^c)$ and the estimated relationship between the prior mathematics score and the outcome ($\hat{\beta}_1$). If the treatment and control groups are balanced on *X* (i.e., they have the same group means) then *X* is not a confounder and the regression-adjusted ATE is the same as the naïve treatment effect estimate. If, however, *X* is a confounder but is omitted from our treatment effect estimate, then that estimate will be biased by $\hat{\beta}_1(\bar{X}^t - \bar{X}^c)$. If we use ANCOVA to adjust for *X*, then our dependence on $\hat{\beta}_1$ for an unbiased estimate of the ATE increases as the mean prior test score difference between treatment and control groups increases (Gelman & Hill, 2006).

The more the treatment and control groups differ in terms of confounders, the more a regression-based approach depends on specifying the correct functional form, or parametric model (Ho et al., 2007; Schafer & Kang, 2008). If, for example, the true relationship between *X* and *Y* is quadratic but we only model the linear relationship depicted in Equation 2.4, then the regression-based treatment effect estimate will be biased by $\beta_2\left(\bar{X}^{2t} - \bar{X}^{2c}\right)$, where β_2 is how the linear relationship between *X* and *Y* changes as *X* increases (Gelman & Hill, 2006). Dependence on the specified parametric model is particularly strong when there is a lack of overlap in the treatment and control covariate distributions (Gelman & Hill, 2006; Ho et al., 2007; King & Zeng, 2005; Schafer & Kang, 2008). If, for example, ten percent of treatment students have prior mathematics test scores higher than any control students, then the estimates of the counterfactual outcome for those treatment students must be extrapolated from the specified parametric model.

Schafer and Kang (2008) warn that treatment effect estimates based on extrapolation make the estimates "unstable and prone to bias" (p. 289) if the regression model is misspecified, and leads to what King and Zeng (2005) call the "dangers of extreme counterfactuals."

Another way in which the utility of a regression-based approach can break down is if heterogeneity in the treatment effect is not properly modeled. Treatment effect heterogeneity implies that the true regression lines for the treatment and control groups are not parallel and, as a result, a standard OLS estimate of β_1 applied to both groups will produce a biased treatment effect estimate (Cochran & Rubin, 1973; Schafer & Kang, 2008). If, for example, the effect of taking an above-grade-level course is larger for students with higher prior mathematics achievement, then one should add a *D*-by-*X* interaction term into the regression-based adjustment model. Morgan and Winship (2007) note that treatment effect heterogeneity is more likely the norm rather than the exception in social science research and warn that "much of the received wisdom on regression modeling breaks down in the presence of individual-level heterogeneity of a causal effect" (p. 142). In the face of treatment effect heterogeneity, an OLS model that fails to model the heterogeneity will estimate a conditional-variance-weighted ATE, where more weight is given to units with *X* values around the treatment group median. This conditional-variance-weighted ATE may not be the researcher's intended estimand of interest (Morgan & Winship, 2007).

In summary, the validity of a regression-based approach to causal effect estimation will depend on three factors: covariate balance, covariate overlap, and the parametric model. If treatment and control groups are perfectly balanced on all confounding covariates, as expected in large-sample randomized experiments, then the regression-adjusted average treatment effect estimate will be the same as an unbiased naïve treatment effect estimate. As the treatment and

control groups diverge from perfect covariate balance, dependence on the assumed parametric model increases. Dependence on the assumed model is especially strong when there is a lack of covariate overlap between treatment and control groups because the researcher must use the model to extrapolate outside the range of the data. The degree to which dependence on the parametric model will bias treatment effect estimates depends on whether the relationship between all the confounding covariates and the outcome are properly modeled and whether the model adequately accounts for treatment effect heterogeneity.

Even if a reasonably specified regression-based approach is employed for treatment effect estimation, the approach alone can cause the researcher to conceptually stray from the intended estimation of causal effects. As Schafer and Kang (2008) clarify, "[t]he ANCOVA treatment effect is an average difference in response *between two different groups of individuals*, adjusting for differences in their covariates. Causal inferences, on the other hand, are about changes in response when different treatments are applied *to the same individuals*" (p. 288, emphasis in original). Additionally, Morgan and Winship (2007) worry that "[f]or causal analysis, the rise of regression led to a focus on equations for outcomes, rather than careful thinking about how the data in hand differ from what would have been generated by the ideal experiments one might wish to have conducted" (p. 13).

2.2.3. Matching-based approaches to causal effect estimation

While regression-based approaches seek to sever the relationship between confounding covariates and the outcome, matching-based approaches seek to sever the relationship between confounding covariates and treatment assignment. Broadly defined, matching can include a variety of methods that all aim to "equate (or 'balance') the distribution of covariates in the

treated and control groups" (Stuart, 2010, p. 1) to remove potential bias in treatment effect estimates that stem from confounding covariates. Matching-based approaches try to create balanced treatment and control groups through either sub-sampling from the treated and control group samples based on observed covariates (Stuart & Rubin, 2007), or weighting the original treatment and control group samples based on observed covariates (Morgan & Harding, 2006). A key characteristic of any matching-based approach is that the method should be implemented without looking at the outcome data (Rubin, 2006) and can therefore be considered a method for preprocessing the data in the design stage before estimation of treatment effects in the analysis stage (Ho et al., 2007).

In its simplest incarnations, researchers invoke matching by restricting group comparisons to units that share similar characteristics. For example, studies of the effects of educational attainment on earnings may restrict the analysis to white males, ensuring that individuals are matched along both race and gender. When researchers wish to control for a single, continuous confounder, such as age, treatment and control units can be matched within subclasses or stratum of the continuous variable (Cochran, 1968). For example, instead of comparing all white males, the comparison can be made within different age group subclasses to help rule out the confounding effect of age on earnings. The feasibility of these rather basic matching approaches often breaks down when one needs to account for a large number of covariates because of data sparseness at any given point in a multivariate joint distribution.

In the 1980s, Rosenbaum and Rubin (Rosenbaum & Rubin, 1983, 1984, 1985) showed that matching on a single scalar summary (or transformation) of multiple covariates, such as the estimated conditional probability of treatment assignment (i.e., the propensity score), can remove bias from all the observed covariates. Their work gave rise to the increasingly common practice

of matching treatment and control units on a predicted propensity score. In fact, the number of articles published in the educational literature that use propensity scores has increased exponentially over the past decade (Thoemmes & Kim, 2011). For example, Attewell and Domina (2008) and Leow, Marcus, Zanutto, and Boruch (2004) used propensity score matching to examine the effects of curricular intensity on academic achievement. Introductions to propensity score matching methods are now common in the social science research literature, including: economics (Caliendo & Kopeinig, 2008; Dehejia & Wahba, 1999), political science (Ho et al., 2007; Sekhon, 2009), psychology (Schafer & Kang, 2008), and sociology (Morgan & Harding, 2006). In her review of matching methods, Stuart (2010) laid out four key steps:

(1) defining the distance measure used for determining matches;

(2) implementing a matching method;

(3) assessing the quality of the resulting matched sample; and

(4) analyzing the outcome based on the matched sample.

The estimated propensity score is the most common distance measure, or balancing score (Rosenbaum & Rubin, 1983), used to define a match when multiple covariates are involved. A balancing score is a function of the observed covariates, *b(X)*, that produces the same covariate distribution in the treated and control groups when conditioning on the balancing score, so that $X \perp D \mid b(X)$. Thus, under the assumption of strongly ignorable treatment assignment, conditioning on the balancing score can produce unbiased average treatment effect estimates. The most precise balancing score is based on the exact values of *X*, while Rosenbaum and Rubin (1983) showed that the propensity score is the coarsest balancing score one can use. Rubin and Thomas (1992a, 1992b) later showed that matching on the estimated propensity score can produce better bias reducing properties than matching on the true propensity score. A

straightforward and common way to estimate the propensity score is with logistic regression, where an individual unit's probability to treatment assignment (p_i) is predicted by the observed confounders (X), so that:

$$\hat{p}_i = \Pr(D_i = 1 \mid X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)},$$

and

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = \beta_0 + \beta_1 X_i.$$

Matching based on the estimated propensity score, or its log-odds linear transformation, can take many forms, from various matching algorithms, to subclassification and weighting (see, for example, Ho et al., 2007; Morgan & Harding, 2006; Schafer & Kang, 2008; Sekhon, 2009; Stuart, 2010). In general, the matching methods "primarily vary in terms of the number of individuals that remain after matching and in the relative weights that different individuals receive" (Stuart, 2010, p. 7). The utility of any propensity score matching method, however, can be assessed by the degree to which covariate balance between the treatment and control groups improved after matching. In other words, the main objective with matching is to create treatment and control groups that have very similar empirical distributions for each pre-treatment covariate, as well as similar multivariate distributions based on interactions between covariates. Distributional comparisons, particularly multivariate distributions, can be difficult to assess when many covariates are included. As a result, balance diagnostics typically focus on univariate comparisons of group means and variance ratios (Rubin, 2001). If an acceptable level of balance is not achieved with a given matching method, the researcher can re-specify the propensity score model (e.g., include non-linear terms) and/or try a different matching method. If balance is achieved through the matching method, then the link between the confounding covariates and

treatment assignment is cut. Referring back to Equation 2.4, having balanced treatment and control groups implies that $\bar{X}^t - \bar{X}^c = 0$ and an unbiased estimate of the average treatment effect can be estimated with the between group mean outcome difference.

The matching-based approach to causal effect estimation has four main advantages over a regression-based approach. First, a matching-based approach is more aligned, conceptually, with the potential outcomes framework and randomized experimentation since the emphasis is on creating a control group in the design phase to provide a counterfactual outcome for the treatment group. Second, since matching is conducted in the design phase, prior to examination of any outcomes, one can repeatedly refine the matching method without concerns of "gaming the analysis" to achieve desired results. Third, by estimating a propensity score and assessing balance, the matching-based approach forces the researcher to consider and examine important factors associated with treatment assignment and the degree of covariate overlap between groups. The process of balance assessment might even reveal a severe lack of overlap that calls into question the feasibility of effect estimation for the entire population, or a specific sub-population. Fourth, by creating treatment and control groups with similar covariate distributions, effect estimation in the analysis phase is less dependent on extrapolation and model-based assumptions regarding the relationship between covariates and the outcomes. Overall, "the advantage of matching is that it is relatively robust to small changes in procedures and produces a data set that is by design less sensitive to modeling assumptions" (Ho et al., 2007, p. 33).

The matching-based approach is not without limitations, however. As with the regression-based approach, strongly ignorable treatment assignment is the key assumption that must hold for unbiased effect estimation. Additionally, while the regression-based approach requires one to properly model the relationship between covariates and the outcome, the

matching-based approach requires one to properly model the relationship between covariates and treatment assignment when estimating the propensity score. As a result, treatment estimates based on a propensity score matched sample may be biased if an unobserved covariate was excluded from the propensity score model or if the functional form relating an observed covariate to treatment assignment was not properly modeled (e.g., omitting a squared term or an interaction term). Problems with the matching-based approach may also arise due to limitations in balance diagnostics for comparing multivariate distributions, or when one must make trade-offs between achieving balance along one set of diagnostic measures versus another set of measures. If one proceeds to the analysis stage without properly determining the extent of covariate balance, the remaining imbalance could bias effect estimates.

Ultimately, even under conditions of strong ignorability, the performance of a matching-based approach depends on the degree of covariate overlap among the original sample and the extent to which the sample contains a large enough reserve of controls to find suitable matches for the treatment units. When the reserve of controls is not large enough—or does not adequately overlap with the treatment group enough—to obtain quality matches for all treatment units, one can focus the analysis on the sub-sample of treatment units with available matches, but must acknowledge that this sub-sample may create "bias due to incomplete matching" and not generalize to the average treatment effect for all treated units (Rosenbaum, 2009). In some settings, it may be possible to expand the reserve of controls by drawing matches from a second control group. For example, in a study of a school dropout prevention program, Stuart and Rubin (2008) first sought matches for treatment students within the program school. For treatment students without a good match within the program school, control students were obtained for a

nearby non-program school. This approach is a key aspect of the proposed methodology and is discussed in more detail in the following chapter.

2.2.4. A note on dual modeling

While the above discussion of regression-based and matching-based approaches to causal effect estimation treats the two approaches as competing methods, they are best used as compliments. As mentioned above, matching-based approaches address the association between background confounders and treatment assignment, and focus on the design stage of the research. Regression-based approaches, on the other hand, address the association between background confounders and the outcome, and focus on the analysis stage of the research. Matching is often proposed as a way to construct a control group in a non-experimental study (Rosenbaum & Rubin, 1985) or preprocess the full sample of data prior to analysis (Ho et al., 2007). Under this conceptualization of matching, one can apply regression-based approaches to the matching-based preprocessed data in the analysis stage to estimate causal effects. Conceptually, this approach parallels the use of ANCOVA in a randomized experiment to adjust for small group differences and increase precision of effect estimates. Taking such an approach in a non-experimental study is often referred to as dual modeling, or doubly robust, because average effect estimates will be unbiased (under the assumption of strong ignorability) if either the propensity score model or the outcome model is properly specified. As such, the dual modeling approach has been found to produce average treatment effect estimates that are less biased and less sensitive to model specification (Robins & Rotnitzky, 1995; Schafer & Kang, 2008).

2.2.5. A note on sensitivity analysis

Regardless of the method for causal effect estimation, when the assignment mechanism is unknown—as in an observational study—one cannot say with certainty that the assumption of strongly ignorable treatment assignment holds. In other words, one can always question whether the estimated treatment effects suffer from selection bias. In their comparison of non-experimental methods, for example, Shadish, Clark, and Steiner (2008) find that adjusting for the correct set of confounders is more important than the adjustment method employed. Sensitivity analysis (Rosenbaum, 2002) allows the researcher to assess whether the substantive conclusions regarding average treatment effects hold when key assumptions like strong ignorability do not hold. For example, Frank (2000) provides a method to determine how correlated an unobserved confounder would have to be with both treatment assignment and the outcome to nullify the finding of a statistically significant average effect. Other approaches to test the sensitivity of results include the use of multiple comparison groups and non-equivalent outcome measures (Shadish et al., 2002). Any non-experimental study should examine whether the findings are sensitive to the key assumptions that support causal effect estimation method.

2.3. Causal effect estimation in a multisite setting

Practices, programs, and policies of interest to social science researchers often occur within a multilevel, or hierarchical, context. This is especially true in education, where, for example, students are nested within classrooms nested within schools. The above discussion of causal inference largely ignores or avoids this multilevel context, yet failure to account for the nested structure of human interaction and program implementation may result in spurious findings and fail to uncover important differences in findings across settings (e.g., across

schools). Extending an analysis to a multilevel setting, however, requires one to address added complications to the framework for causal inference. In this section, I discuss how moving from a single-site study to a multisite study complicates the assumptions of strongly ignorable treatment assignment and SUTVA, and how some researchers have addressed these complications.

2.3.1. The strongly ignorable treatment assignment assumption in a multisite setting

For the strongly ignorable treatment assignment assumption to hold, the assignment mechanism must be independent of the potential outcomes conditional on the observed covariates (recall panel B in Figure 2.1). In a single site study, the observed individual-level pretreatment characteristics (X) might allow a researcher to properly condition on the observables to invoke the ignorability assumption. In a multisite study, however, it is possible that site-level characteristics influence both treatment assignment and outcomes. Thus, ignoring site-level covariates would invalidate the ignorability assumption. For example, schools with high academic expectations for their students might encourage students to take the above-grade-level course and might also place greater emphasis on academic achievement compared to schools with lower academic expectations. Therefore, student selection into an above-grade-level course will depend on both student-level characteristics and the student's school.

The types of assignment mechanisms depicted in Figure 2.1 can be amended to include school-level factors, S, that can influence treatment assignment under a multisite study. In Figure 2.2, panel A represents a situation where both school assignment and treatment assignment are independent of student-level observables (X). Such a situation might arise if students were randomly assigned to schools and treatment conditions, and an analysis that ignored both student

and school factors could produce unbiased treatment effect estimates. Panel B depicts a situation in which both school assignment and treatment assignment depends on observed student-level characteristics, but school-level characteristics are only associated with treatment assignment through *X*. For example, students are partially sorted in schools based on prior achievement and prior achievement influences treatment assignment. Under such a condition, one could produce unbiased treatment effect estimates by conditioning on *X* but ignoring *S*. In Panel C, however, treatment assignment depends on both student-level and school-level observed characteristics and ignoring these characteristics in the analysis would bias the results. Rumberger (1995), for example, studied the factors associated with dropping out of school and found that both student and school-level characteristics significantly predicted dropping out. Given this finding, studies that seek to estimate causal effects of dropping out would have to control for both student and school characteristics, assuming they are associated with the outcome(s) of interest.

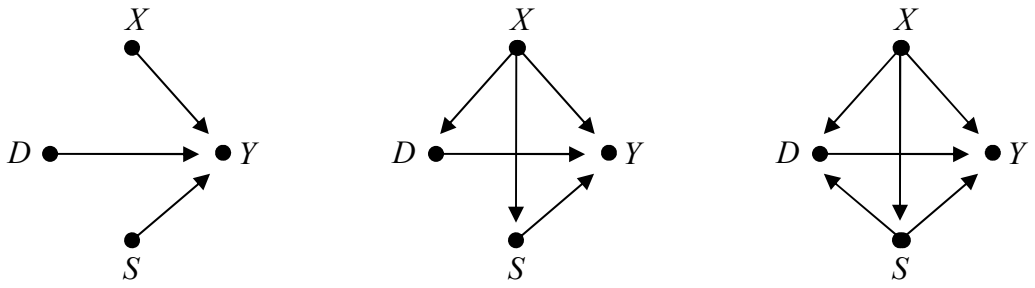

Figure 2.2. Three different types of assignment mechanisms that can arise in a multisite study.

Another potential complication under a multisite setting is that the student-level characteristics necessary to validate the ignorability assumption could differ across sites. For

example, assignment to an above-grade-level course might depend on prior achievement in one school, but might depend on prior achievement and an entrance exam in another school. Furthermore, school-level characteristics might interact with student-level characteristics in the determination of treatment assignment and/or outcomes, thus invalidating treatment effect estimates that condition on observed student- and school-level characteristics but fail to account for the interactions. Kim and Seltzer (2007), for example, found that high school student participation in an academic outreach program aimed at increasing college eligibility significantly differed across schools both because of overall differences in school participation rates and variation in the predictive importance of student-level factors across schools. Similarly, Rickles (2011) studied the process by which 8th graders are assigned to algebra or pre-algebra and found school-level variation in the weight placed on different student factors. Note that these complications are not easily represented in graphical diagrams like Figure 2.2.

2.3.2. The SUTVA assumption in a multisite setting

SUTVA is comprised of two assumptions that allow one to simplify the definition of a causal effect by restricting the number of potential outcomes for any given treatment. The first component of SUTVA states that the potential outcomes for any given unit do not depend on the treatment assignment of another unit. The second component states that a given treatment is stable across units. If both SUTVA components hold, one can define the causal effect between two treatment conditions for a given unit based on two potential outcomes: one under treatment, $Y(1)_i$, and one under control, $Y(0)_i$. However, both SUTVA components are not likely to hold in a multisite setting where the potential outcomes for any one unit could depend on group membership and result in what Gitelman (2005) called an "infinite collection [of potential

outcomes] for each subject (the 'fundamental problem of causal inference' in the extreme)" (p. 404).

The assumption that the potential outcomes of one unit are not affected by the assignment of another unit is more commonly discussed as an assumption of no interference between units (Raudenbush, 2008; Sobel, 2007). Within an educational multilevel setting, this assumption can break down from peer effects within treatment groups or spillover effects arising from units interacting between groups. Under such conditions, the potential outcome for a given student depends not only on the treatment received but also on the interaction with other students. Consider, for example, the potential outcomes for a high achieving student, i=1, and the potential outcomes for a low achieving student, i=2, if assigned to an above-grade-level mathematics course (D=1) or the on-grade-level mathematics course (D=0). Under SUTVA, each student only has two potential outcomes. For student 1, the potential outcomes are $Y_{i=1}(D_{i=1} = 1)$ and $Y_{i=1}(D_{i=1} = 0)$ regardless of whether student 2 is assigned to the above-grade-level or on-grade-level course. Given peer effects and just two students, however, student 1 has four potential outcomes that depend not just on what course student 1 is assigned to, but also on what course student 2 is assigned (see Table 2.1). As the number of students with possible interaction with student 1 increases, the number of potential outcomes for student 1 increases geometrically and makes meaningful causal effects difficult to define.

Table 2.1. Potential outcomes for a single unit when there is interference from another unit.

	$D_{i=2} = 1$	$D_{i=2} = 0$
$D_{i=1} = 1$	$Y_{i=1}(D_{i=1} = 1, D_{i=2} = 1)$	$Y_{i=1}(D_{i=1} = 1, D_{i=2} = 0)$
$D_{i=0} = 0$	$Y_{i=1}(D_{i=1} = 0, D_{i=2} = 1)$	$Y_{i=1}(D_{i=1} = 0, D_{i=2} = 0)$

Similarly, the assumption that a treatment is stable across units implies that the treatments do not vary across groups. In education, this assumption may hold for fairly prescribed interventions that are easily implemented, but likely breaks down for any treatment aimed at altering the instructional environment. Given variations in instructional practices and quality across schools and classrooms, a treatment such as assignment to an above-grade-level course could mean different things depending on the teachers, peers, and instructional materials associated with the course in a school or classroom. In other words, a student's potential outcomes will not depend on the instructional environment within which the treatment is implemented. Hong and Raudenbush (2006) refer to this type of organizational effect as treatment enactment variation. As with the above complication of interference, treatment enactment variation can result in an unmanageable number of potential outcomes for a given unit. To see this, again consider student 1, who could be assigned to an above-grade-level or on-grade-level course. Unlike the above situation, however, assume no interference but two types of classroom environments for each treatment condition: one with high quality instruction (Q=1) and one without high quality instruction (Q=0). Under this scenario, student 1 has four potential outcomes (see Table 2.2) instead of two and defining a causal effect is complicated.

Table 2.2. Potential outcomes for a single unit when there is treatment enactment variation.

	$Q = 1$	$Q = 0$
$D_{i=1} = 1$	$Y_{i=1}(D_{i=1} = 1, Q = 1)$	$Y_{i=1}(D_{i=1} = 1, Q = 0)$
$D_{i=0} = 0$	$Y_{i=1}(D_{i=1} = 0, Q = 1)$	$Y_{i=1}(D_{i=1} = 0, Q = 0)$

When SUTVA does not hold under a multilevel setting, some researchers have addressed the ambiguity in causal effect definitions by redefining the level of treatment and/or relaxing the

SUTVA components. A common approach in experimental studies is to conduct a cluster randomized design (Raudenbush, 1997), where intact groups (e.g., classrooms or schools) are assigned to treatments instead of individuals. Under this design, the possibility of peer and spillover effects is limited because individuals are much less likely to associate between groups than within groups. Raudenbush (2008) discusses two assumptions embedded in this design: (1) that there is no interference between groups and (2) groups remain intact. In other words, individual interactions between groups do not influence potential outcomes and individuals do not change groups during the treatment period. Similarly, Gitelman (2005) discusses using an assumption of group-membership invariance—which maintains an assumption of no group-dynamic effect (e.g. no peer effects) but allows for an individual's potential outcomes to vary based on group-level characteristics—to define a "group-allocation, multilevel average" (GAMA) casual effect. Hong and Raudenbush (2005, 2006) utilize the "no interference between groups" assumption to estimate the effect of kindergarten retention in a non-experimental study. They do, however, allow within-group peer effects to work through a scalar function that reduces the number of potential outcomes to a manageable set. In their case, they assumed that peer effects could influence an individual student's potential outcomes by either being exposed to a high proportion of retained peers or not. Hong and Raudenbush (2008) also extended this approach to estimate causal effects of time-varying instructional treatments.

2.3.3. Randomized experiments in a multisite setting

As when designing a single-site observational study, it is important to follow Dorn's advice and first consider how the study would be conducted if it were possible to do a randomized experiment. In a multisite setting, two main randomized designs are commonly

employed, with the main distinction being the level at which treatments are assigned. One design, briefly discussed above, is the cluster randomized design (Raudenbush, 1997). With this design, treatments are randomly assigned at the group-level instead of the individual unit level and treatment effects are estimated by comparing the treatment and control group-level outcomes. The other design is the multisite randomized design, or block randomized design (Seltzer, 2004). With this design, within each group (or block), treatments are randomly assigned at the individual unit level and treatment effects are estimated by comparing the treatment and control unit-level outcomes within each group. For example, within each school in a study, students are randomly assigned to a treatment or control condition. The multisite randomized design was employed, for example, by Project STAR to study the effects of class size reduction (Finn & Achilles, 1990). In the study, students were randomly assigned to classes within each of the 76 participating schools. Unlike the cluster randomized design, the multisite randomized design allows for treatment effect estimates within each site. As such, one can treat the sites as separate mini-experiments and use meta-analytic concepts to investigate reasons for treatment effect variation across the sites (Seltzer, 2004).

2.3.4. Regression-based approaches to causal effect estimation in a multisite setting

In non-experimental studies, regression-based methods to estimate causal effects are commonly extended to address multisite settings with multilevel or hierarchical models (Raudenbush & Bryk, 2002). Hierarchical models (HM) have been used, for example, to study the effects of high school tracking (Gamoran, 1992) and curricular intensity (Wang & Goldschmidt, 2003) on mathematics achievement. HMs have three key characteristics that make them useful for estimating causal effects in a multisite setting. First, it provides a mechanism for

conditioning on both unit- and group-level covariates, as well as any cross-level interactions between covariates. This conditioning is necessary to identify unbiased effect estimates when both types of factors influence treatment assignment and the outcome(s) of interest (e.g., Figure 2.2, panel c). Second, it allows one to model and test for any site-level heterogeneity in treatment effects, which allows one to conceptually parallel the multisite randomized design discussed above. Third, given site-level treatment effect heterogeneity, one can use a HM to examine what site-level factors are associated with effect size.

As an example, consider the above-grade-level course treatment where we are now interested in the effect on students (indexed by *i*) across multiple schools (indexed by *j*). One could set up the following two-level HM with random intercept and slopes to estimate the treatment effect:

$$\begin{aligned}
&\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij}^{gp} + \delta_j D_{ij} + e_{ij}, && e_{ij} \sim N(0, \sigma^2), \\
&\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01} S_j^{gd} + u_{0j}, && \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{\delta j} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \tau_0 & . & . \\ \tau_{01} & \tau_1 & . \\ \tau_{0\delta} & \tau_{1\delta} & \tau_\delta \end{bmatrix} \right). \\
&\qquad \beta_{1j} = \gamma_{10} + \gamma_{11} S_j^{gd} + u_{1j}, && \\
&\qquad \delta_j = \gamma_{\delta 0} + \gamma_{\delta 1} S_j^{gd} + u_{\delta j}, &&
\end{aligned} \tag{2.5}$$

Or the two-level model can be expressed in combined mixed-model form:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij}^{gp} + \gamma_{\delta 0} D_{ij} + \gamma_{01} S_j^{gd} + \gamma_{11} (S_j^{gd}) X_{ij}^{gp} + \gamma_{\delta 1} (S_j^{gd}) D_{ij} + u_{1j} X_{ij}^{gp} + u_{\delta j} D_{ij} + u_{0j} + e_{ij}.$$

With this model, we can adjust for differences in student-level prior achievement, X_{ij}^{gp}, where the superscript *gp* indicates that the prior achievement score is centered on the group mean, and differences in a school-level factor, S_j^{gd}, where the superscript *gd* indicates that it is centered on the grand mean. With this centering, β_{0j} is the score for an average control student in school *j* and δ_j is the treatment effect for the average student in school *j*. Similarly, γ_{00} is the overall average control student score and $\gamma_{\delta 0}$ is the overall average treatment effect, or GAMA

(Gitelman, 2005). In this model, the average treatment effect for a given school depends on a cross-level interaction ($\gamma_{\delta 1}$), or contextual effect, based on *S* and residual school-level deviations from the overall average treatment effect, $u_{\delta j}$. The variance parameter τ_{δ} represents the degree of between-school heterogeneity in the treatment effect after accounting for *S*.

From the mixed-model, we can rearrange the terms to define the observed outcome for a given unit in a sample as deviations from the grand-mean (γ_{00}) based on overall and site-specific adjustments for *X*, *D*, and *S*, plus unique site- and unit-specific effects:

$$Y_{ij} = \gamma_{00} + (\gamma_{10} + u_{1j})X_{ij}^{gp} + (\gamma_{\delta 0} + u_{\delta j})D_{ij} + (\gamma_{01} + \gamma_{11}X_{ij}^{gp} + \gamma_{\delta 1}D_{ij})S_j^{gd} + u_{0j} + e_{ij}$$

Rearranging the terms facilitates a definition for the observed treatment group mean outcome based on the following formula:

$$\overline{Y}^t = \hat{\gamma}_{00} + \hat{\overline{u}}_0^t + (\hat{\gamma}_{10} + \hat{\overline{u}}_1^t)(\overline{X}^t - \overline{X}) + (\hat{\gamma}_{01} + \hat{\gamma}_{11}(\overline{X}^t - \overline{X}) + \hat{\gamma}_{\delta 1})(\overline{S}^t - \overline{S}) + (\hat{\gamma}_{\delta 0} + \hat{\overline{u}}_{\delta}^t).$$

Similarly, the observed control group mean outcome can be defined as:

$$\overline{Y}^c = \hat{\gamma}_{00} + \hat{\overline{u}}_0^c + (\hat{\gamma}_{10} + \hat{\overline{u}}_1^c)(\overline{X}^c - \overline{X}) + (\hat{\gamma}_{01} + \hat{\gamma}_{11}(\overline{X}^c - \overline{X}))(\overline{S}^c - \overline{S}).$$

The above quantities now have the following interpretations:

- $\hat{\gamma}_{00}$ = estimated grand-mean outcome for control students;
- $\hat{\overline{u}}_0^t$ = estimated treatment group mean school residual deviation from the predicted grand-mean outcome for control students;
- $\hat{\overline{u}}_0^c$ = estimated control group mean school residual deviation from the predicted grand-mean outcome for control students;
- $\hat{\gamma}_{10}$ = estimated grand-mean slope relating changes in *X* to changes in *Y*;

- $\hat{\bar{u}}_1^t$ = estimated treatment group mean school residual deviation from the predicted grand-mean slope relating changes in *X* to changes in *Y*;
- $\hat{\bar{u}}_1^c$ = estimated control group mean school residual deviation from the predicted grand-mean slope for *X* on *Y*;
- $\hat{\gamma}_{\delta 0}$ = estimated grand-mean treatment effect
- $\hat{\bar{u}}_\delta^t$ = estimated treatment group mean school residual deviation from the predicted grand-mean treatment effect;
- $\hat{\gamma}_{01}$ = estimated slope relating changes in *S* to changes in the estimated mean outcome for control students;
- $\hat{\gamma}_{11}$ = estimated slope relating changes in *S* to changes in the estimated slope for *X* on *Y*;
- $\hat{\gamma}_{\delta 1}$ = estimated slope relating changes in *S* to changes in the estimated treatment effect;

As with an OLS regression (see Equation 2.4), the above HM-based definitions for the observed treatment and control group means, allow one to define the HM-adjusted average treatment effect ($\hat{\gamma}_{\delta 0}$) as an adjustment from the observed group mean difference ($\bar{Y}^t - \bar{Y}^c$) as:

$$\begin{aligned}\hat{\gamma}_{\delta 0} = \bar{Y}^t - \bar{Y}^c - (\hat{\gamma}_{10} + (\hat{\bar{u}}_1^t - \hat{\bar{u}}_1^c))(\bar{X}^t - \bar{X}^c) \\ - \ (\hat{\gamma}_{01} + \hat{\gamma}_{11}(\bar{X}^t - \bar{X}^c) + \hat{\gamma}_{\delta 1})(\bar{S}^t - \bar{S}^c) - (\hat{\bar{u}}_0^t - \hat{\bar{u}}_0^c) - \hat{\bar{u}}_\delta^t.\end{aligned} \tag{2.6}$$

While Equation 2.6 looks convoluted, it can be broken into five components that facilitate a better understanding for how the HM adjusts the naïve treatment effect for different sources of bias. These components are discussed below:

- $\bar{Y}^t - \bar{Y}^c$ = the naïve treatment effect (e.g., unadjusted mean group difference).
- $(\hat{\gamma}_{10} + (\hat{\bar{u}}_1^t - \hat{\bar{u}}_1^c))(\bar{X}^t - \bar{X}^c)$ = adjustment for differences in *X*, where the magnitude of the adjustment consists of an overall adjustment factor, $\hat{\gamma}_{10}$, and an additional adjustment

arising from between-school heterogeneity in the relationship between *X* and *Y* when treatment and control groups are not evenly distributed across schools ($\hat{\bar{u}}_1^t - \hat{\bar{u}}_1^c$). If there is no between-school heterogeneity in the *X*-*Y* relationship after adjusting for *S*, or the residual school-level heterogeneity in the relationship is not associated with treatment assignment, then the $\hat{\bar{u}}_1^t - \hat{\bar{u}}_1^c$ term drops out of the adjustment. Furthermore, if treatment and control groups have the same mean for *X*, or *X* is not related to *Y*, then the entire component reduces to zero.

- $(\hat{\gamma}_{01} + \hat{\gamma}_{11}(\bar{X}^t - \bar{X}^c) + \hat{\gamma}_{\delta 1})(\bar{S}^t - \bar{S}^c)$ = adjustment for differences in *S*, where the magnitude of the adjustment consists of an overall adjustment factor, $\hat{\gamma}_{01}$, an adjustment for the interaction of *S* with *X*, $\hat{\gamma}_{11}(\bar{X}^t - \bar{X}^c)$, and an adjustment for heterogeneity in the treatment effect due to *S*, $\hat{\gamma}_{\delta 1}$. If S is not related to Y, then the $\hat{\gamma}_{01}$ term drops out of the adjustment. If the *X*-*Y* relationship does not depend on values of S or the two groups have the same mean for *X*, then the $\hat{\gamma}_{11}(\bar{X}^t - \bar{X}^c)$ term drops out of the adjustment. Similarly, if the treatment effect does not depend on values of *S*, then the $\hat{\gamma}_{\delta 1}$ term drops out of the adjustment. Furthermore, if treatment and control groups have the same mean for *S*, which would be guaranteed if the two groups are evenly distributed across schools, then the entire component reduces to zero.
- $(\hat{\bar{u}}_0^t - \hat{\bar{u}}_0^c)$ = adjustment for between-school heterogeneity in control group means after adjusting for *S*. If the treatment and control groups are evenly distributed across schools, or the residual school-level heterogeneity in control group means is not related to treatment assignment, then this component reduces to zero.

- $\hat{\bar{u}}_{\delta}^{t}$ = adjustment for between-school heterogeneity in the treatment effect after adjusting for *S*. If the treatment and control groups are evenly distributed across schools, or the residual school-level heterogeneity in the treatment effect is not related to treatment assignment, then this component reduces to zero.

As with regression-based effect estimation in the single-site example, validity of the regression-based effect estimates in the multilevel setting depends on parametric assumptions about how the estimated parameters reflect the true associations between the outcome and covariates, and one's dependence on those parametric assumptions increases with the size of the initial difference(s) between treatment and control groups. Given the need to estimate site-level parameters and site-level residuals in the multilevel setting, the dependence on parametric assumptions can be even greater than in the single-site setting. It is important to recognize, however, that the relative degree of dependence on parametric assumptions is based on the initial treatment-control group differences and the study design. For example, if the treatment assignment mechanism depicted in Figure 2.2c holds, then adjusting for differences in *X* and *S* are sufficient to recover an unbiased treatment effect estimate. In this case, estimates of the average school residual terms (i.e., the $\hat{\bar{u}}'s$) will converge to zero and not factor into the adjustment. Furthermore, if the assignment mechanism does not depend on school-level factors, so treatment and control students are evenly distributed across schools, then all the adjustment components involving either *S* and/or the $\hat{\bar{u}}'s$ drop out of the adjustment. In this case, one is left with $\hat{\gamma}_{\delta 0} = \bar{Y}^{t} - \bar{Y}^{c} - \hat{\gamma}_{10}(\bar{X}^{t} - \bar{X}^{c})$, which is identical to the single-site adjustment formula in Equation 2.4.

Additionally, one can manipulate Equation 2.6 to define HM-adjusted within-school treatment effects, which is desired from a multisite study conceptualized as a multisite randomized block design. Under this design, it is not necessary to include school-level covariates (e.g., *S*) since treatment and control students are compared within the same schools and will, by definition, not differ along school-level covariates. Therefore, a HM that excludes school-level covariates would estimate an adjusted treatment effect for each school defined as:

$$(\hat{\gamma}_{\delta 0} + \hat{u}_{\delta j}) = \overline{Y}^{t} - \overline{Y}^{c} - \hat{u}_{0j} - (\hat{\gamma}_{10} + \hat{u}_{1j})(\overline{X}_{j}^{t} - \overline{X}_{j}^{c}), \tag{2.7}$$

where $\hat{u}_{\delta j}$ indicates how the estimated treatment effect in school *j* differs from the grand-mean treatment effect, and the HM-adjustment from the overall naïve treatment effect allows for variation in the school-specific control group average ($\hat{u}_{0j}$) and slope relating *X* to *Y* ($\hat{u}_{1j}$). Conducting a regression-based adjustment that ignores the multilevel structure of the data and possible school heterogeneity would estimate a treatment effect that is biased for school *j* by the amount equal to $\hat{u}_{0j} + \hat{u}_{1j}(\overline{X}_{j}^{t} - \overline{X}_{j}^{c}) - \hat{u}_{\delta j}$.

In this subsection, I showed how the use of a HM can produce regression-based average treatment effect estimates that adjust for student- and school-level covariates, as well as produce school-specific treatment effect estimates that allow one to investigate school-level heterogeneity in treatment effects. This second use for a HM mirrors the utility of a multisite randomized design where schools are the blocking factor. By breaking down how the HM-adjusted average treatment effect differs from the observed naïve treatment effect, however, one can see how these regression-based adjustments could be sensitive to the parametric model assumptions. This is particularly true when initial treatment and control group differences are substantial. In the next subsection I discuss the matching-based approach to causal effect estimation, which focuses on ways to minimize these group differences.

2.3.5. Matching-based approaches to causal effect estimation in a multisite setting

Relative to the use of regression-based HMs for causal effect estimation in multisite settings, matching-based approaches are rarely used and the methodology is in its infancy. Matching-based approaches, however, are a promising way to reduce a non-experimental multisite study's dependence on modeling assumptions, as well as better align the estimation of causal effects with the potential outcomes framework and the conceptualization of a multisite randomized experiment. In discussing the issues that complicate matching-based approaches in a multisite setting, I follow the four key steps Stuart (2010) outlined for matching methods, as discussed above for single-site studies.

The first step for a matching-based approach is to define the distance measure one will use to determine matches. In a single-site setting, the most common distance measure is a propensity score estimated from a logistic regression model that predicts each individual unit's propensity for treatment assignment given the unit's covariate values. In a multisite setting, one must consider whether site-level covariates should be included in the model for the assumption of strong ignorability to hold. Furthermore, one must consider whether types of covariates, and the relative importance of covariates, differ across sites. If, for example, prior achievement is an important predictor of above-grade-level course placement in one school, but not another, using a propensity score model based on one of the schools may not produce well matched groups for the other school. Similarly, a simple propensity score model based on the pooled sample from both schools many not have adequate balancing properties for either school. Depending on the degree of between-site heterogeneity in the assignment mechanism, one may want to construct a separate propensity score model for each site. Within-site sample sizes, however, may make

estimation of separate site-level propensity scores unstable, particularly when a large number of covariates are needed.

Another option is to estimate each unit's propensity score from a logistic hierarchical regression model (LHM), which can incorporate random intercepts (i.e., allow for site-level variation in the average propensity score) and random slopes (i.e., allow for site-level variation in the importance of different covariates) (Kim & Seltzer, 2007). Hong and Raudenbush (2005), for example, employed a LHM with a random intercept to estimate the propensity score for kindergarten retention. More generally, a LHM with random intercept and slopes could take the following form for estimating the log-odds of treatment assignment for student *i* nested within school *j* (D_{ij}=1):

$$\begin{aligned} &\text{Level 1: } \ln\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = \beta_{0j} + \beta_{1j} X_{ij}^{gp}, \\ &\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01} S_j^{gd} + u_{0j}, \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \tau_0 & . \\ \tau_{01} & \tau_1 \end{bmatrix}\right), \\ &\qquad\qquad \beta_{1j} = \gamma_{10} + \gamma_{11} S_j^{gd} + u_{1j}, \end{aligned} \tag{2.8}$$

where $\hat{p}_{ij}$ is the estimated probability of treatment assignment for a given student. The Level 2 model for β_{0j} allows the average propensity score for a student to vary based on a school-level covariate, *S*, and a residual school-level random effect, u_{0j}. Similarly, the Level 2 model for β_{1j} allows the slope for *X* on *Y* to vary based on the school-level covariate and a residual school-level random effect, u_{1j}.

A handful of studies investigated the performance of multilevel propensity score models and generally conclude that, within a multilevel setting, ignoring the multilevel data structure in propensity score estimation will produce more biased treatment effect estimates than when the propensity score model incorporates the hierarchical structure. Kim and Seltzer (2007)—in a

within-school matching analysis of an academic outreach program—demonstrated that when the effects of Level 1 covariates differ across schools, matching based on a model that allows β_0 but not β_1 to vary across schools will produce poor matches compared to a model that allows both β_0 and β_1 to vary across schools. Similarly, Thoemmes and West (2011) conducted simulations to compare the performance of different multilevel propensity score models (random intercept vs. random intercept and one random slope vs. random intercept and two random slopes) to a single-level model. They found little bias when effect estimates were based on a multilevel propensity score model, but significant bias from a single-level propensity score model, even when the important covariates were included in the model. The Thoemmes and West simulation results further showed that, while bias was small for each of the different multilevel propensity score model specifications, bias was greater when slopes were not allowed to vary across groups. Two other simulation studies (Arpino & Mealli, 2011; Su & Cortina, 2009) concluded that ignoring the group level in propensity score estimation will result in biased estimates. One interesting finding from Arpino and Mealli (2011) was that a fixed-effects propensity score model performed better than a random-effects model, but they did not investigate conditions under which slopes could vary across groups.

After deciding on a distance measure—likely from a propensity score model that could ignore site-level differences, include fixed-effects, or allow the intercept and/or slopes to vary across sites—the next step is to implement a matching method. Given the matching choices one must make in a single-site setting (see above discussion in 2.2.3), the primary decision in a multisite setting is whether available matches should be restricted to units within the same site or whether units can be matched across sites. Within-site matching is generally preferred, primarily because matching a treatment student to a control student within the same school "controls all

observed and unobserved pretreatment variables that are constant for all students within a school (e.g., average per-pupil expenditures) and, moreover, provides good control for geographic variables (e.g., urban vs. suburban vs. rural residence)" (Rosenbaum, 1986, p. 210). Returning to the formula for a regression-based adjustment of the naïve treatment effect (Equation 2.6), within-school matching means that not only will pre-treatment group differences in *S* drop out of the adjustment, but so will any residual (i.e., unobserved) school-level differences. As a result, one only needs to worry about adjustments for Level 1 group differences. Within-site matching has the further benefit that it preserves the randomized block design conceptualization by maintaining comparisons within the same block (e.g., school).

Within-site matching may not be feasible, however, because of within-school sample size constraints and/or limited within-school covariate overlap (Kelcey, 2011a). Studies, for example, based on large-scale national datasets typically only contain a small sample of units within sites and there may not be enough control units for each treatment unit within a given site. More generally, given significant pre-treatment group differences along important covariates, one may not be able to find "good" matches for all the treatment units within any given site. The lack of good within-site matches may even be true with relatively large within-site sample sizes depending on the degree of propensity score overlap among treatment and control groups.

As a result, it may be necessary to allow for between-site matching rather than lose a large number of unmatchable treatment units. Between-site matching may work well if one is not concerned about unobserved site-level covariates and the propensity score properly incorporates the site-level covariates, because matching can balance both the level-1 and level-2 covariates. For example, in their study of kindergarten retention, Hong and Raudenbush (2005, 2006) estimated each student's propensity score based on student-level covariates, school-level

covariates, and a school random effect, then use propensity score subclassification for effect estimation, where students in different schools could be in the same subclass. If, however, one is concerned about unobserved site-level covariates, matching between-sites could produce more biased results than if one could match within-sites.

In practice, the decision to use within-site matching or between-site matching can be thought of as a decision to prioritize balance among Level-1 covariates or balance among Level-2 covariates. Restricting matches to the same site will ensure balance among Level-2 covariates, but may require matching some treatment units to control units who do not have very similar Level-1 covariates. The likelihood of finding treatment and control units with similar Level-1 covariates should increase, however, if between-site matching is allowed, but then balance among Level-2 covariates is not guaranteed. It is interesting to note that the discussion of within- and between-site matching aligns with Campbell's emphasis to have a focal local (Shadish & Cook, 2009) control group, where "focal" refers to the Level-1 covariates and "local" refers to the Level-2 covariates. Ideally, one wants matches that are both focal and local, but it may be necessary to prioritize one type of match over another.

One possible compromise for the within- and between-site matching decision is to conduct the matching in stages. Stuart and Rubin (2008) outlined a two-stage matching process they applied to the analysis of a school dropout prevention program where there was not enough covariate overlap, nor enough potential control matches, to match students within the same district. In the first stage, treatment students are matched to local control students and matches that fall within an acceptable propensity score caliper range (e.g., within 0.25 of a standard deviation) are kept. In the second stage, treatment students without an acceptable match from stage one are matched to non-local control students. Stuart and Rubin also provide a method to

adjust the non-local matches from the second stage for the possible introduction of site-level bias. A key component of the proposed method for estimating treatment effects in a multisite setting is the extension of Stuart and Rubin's two-stage matching process to a study with multiple treatment sites. Both the Stuart and Rubin approach and my proposed method are described in detail in the next chapter.

The choice of within- or between-school matching, or even a two-stage matching process, is further complicated by the fact that the balancing potential of a match could depend on the type of distance measure (e.g. propensity score model) employed. For example, the choice to use a single-level, fixed-effects, or random-intercept propensity score model is moot if one plans to conduct within-site matching. This is because the fixed-effect and random-intercept model only provide a uniform adjustment to the estimated propensity score based on site, so the relative ranking of units within a given site will not change from one model to the next. Within-site matching with a propensity score model that includes random-slopes, however, could result in different matches—or even no matches if a caliper range is employed—compared to the other propensity score models (Kim & Seltzer, 2007). Furthermore, when conducting between-site matching based on an estimated propensity score model with random-slopes, units with similar estimated propensity scores in two different sites may not share similar covariate values (Kelcey, 2011b; Steiner, 2011).

The complications regarding propensity score model selection and the matching method means the third step for implementing a matching method is especially important: assessing the quality of the resulting matched sample. As with a single-site study, one wants the matched treatment and control groups to exhibit similar covariate distributions and the standardized mean difference and variance ratio provide key summaries of the covariate overlap. In a multisite

study, where the desire is to make inferences about site-level treatment effects, one must not only assess the quality of the matches across the entire sample, but must also assess the matching quality within each site. If acceptable balance is not achieved, one can try to re-specify the propensity score model and/or try a different matching method.

When acceptable covariate balance is achieved through matching, the final step is to analyze the outcome based on the matched sample. In a multisite setting, the main decision in the analysis stage is how to summarize the treatment effect within and between the sites (Rubin, 1981). One option is to separately estimate an average treatment effect within each site. For example, one could simply estimate the group mean difference within each site, or employ an OLS model within each site to adjust for any covariate differences that remain after matching. The average within-site treatment effects could then be averaged to estimate an overall average treatment effect (i.e., the GAMA). Another option is to take a more meta-analytic approach and treat each site as a mini-study within which the within-site average treatment effect reflects both true site-effect variability and sampling variability from the overall average treatment effect. From this perspective, one could use a multilevel model to estimate the GAMA and site-specific treatment effects based on empirical Bayes, or shrinkage, estimates. A multilevel model was paired with propensity score subclassification by Hong and Raudenbush (2005), for example, to examine variation in the effect of kindergarten retention across schools.

In general, the average treatment effects can be estimated from the following model:

$$\begin{aligned} &\text{Level 1: } Y_{ij} = \beta_{0j} + \delta_j D_{ij} + e_{ij}, \qquad e_{ij} \sim N(0, \sigma^2), \\ &\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \qquad \begin{bmatrix} u_{0j} \\ u_{\delta j} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \tau_0 & . \\ \tau_{0\delta} & \tau_\delta \end{bmatrix}\right), \\ &\qquad\qquad \delta_j = \gamma_{\delta 0} + u_{\delta j}, \end{aligned} \tag{2.9}$$

where $\hat{\gamma}_{\delta 0}$ represents the GAMA and $\hat{\delta}_j$ is the empirical Bayes average treatment effect for site j. The empirical Bayes estimate is a weighted average of $\gamma_{\delta 0}$ and the observed average treatment effect for site *j*, where the weight is based on the precision of the within-site estimate:

$$\hat{\delta}_j = \left(1 - \frac{\hat{\tau}_\delta}{\hat{\tau}_\delta + \hat{v}_{\delta j}}\right)\hat{\gamma}_{\delta 0} + \left(\frac{\hat{\tau}_\delta}{\hat{\tau}_\delta + \hat{v}_{\delta j}}\right)\hat{\delta}_j^{obs}.$$

For the weight, $\hat{v}_{\delta j}$ is the estimated error variance for the average treatment effect estimate in site *j*, which decreases as sample size for school *j* increases (Raudenbush & Bryk, 2002; Rubin, 1981). As with single-site estimation, one can incorporate covariates into the above model to adjust for any pre-treatment group differences that remained after matching.

The empirical Bayes estimated site-level average treatment effect can improve estimation of site-specific treatment effects because it borrows information from other sites. Observed within-site average effects are shrunk toward the GAMA based on the degree of within-site error variance. Including important site-level characteristics at level 2 will shrink the emperical Bayes estimates toward a conditional grand-mean instead of an overall mean, which can improve plausibility in the assumption about site-level exchangeability. Rubin (1981) demonstrated the utility of this approach, and a more general Bayesian approach, with an analysis of a SAT coaching program conducted in multiple schools. Additionally, Su and Cortina (2009) found, through simulations, that the combination of propensity score matching with multilevel modeling produced less biased causal estimates than either directly modeling outcomes or using a single-level model after propensity score matching.

2.4. Summary of Concepts and Contribution of the Proposed Method

In this chapter I reviewed the general conceptualization of causal effect estimation within the potential outcomes framework, along with the main design and analysis considerations one must address when trying to estimate causal effects. The proposed method aims to facilitate causal effect estimation in a multisite setting where one expects heterogeneity in both treatment assignment and treatment effects, and is grounded in two main ideas that come out of the above review.

First, in a non-experimental study it is useful to follow Dorn's advice (Cochran, 1965) and consider how one would like to design the study if a randomized experiment was possible. In a multisite setting where one wants to draw inferences about the heterogeneity of treatment effects across sites, a multisite randomized design with sites as the blocking factor would be an ideal design. As a result, the proposed method seeks to design the study in a way that mimics the multisite randomized design by constructing within-site comparison groups that allow for site-specific treatment effect estimates. A key aspect of this approach is imposing a distinction between the design stage and analysis stage of a study (Rubin, 2001), and using matching-based methods as a data preprocessing tool (Ho et al., 2007) to construct site-specific comparison groups. When trying to match within sites, sample size and limited covariate overlap can be a complication. Therefore, the proposed method adapts the two-stage matching approach developed by Stuart and Rubin (2008) to a multisite design to try and maximize the extent to which the comparison group is both focal and local (Shadish & Cook, 2009).

Second, multilevel regression models provide a way to capture both the between-site heterogeneity in the treatment assignment mechanism and treatment effects. In the design stage, a logistic multilevel model with random intercept and slopes can be employed to estimate

propensity scores for units nested within sites (Kim & Seltzer, 2007); an approach the proposed method employs. In the analysis stage, both site-specific and the overall average treatment effects can be estimated efficiently with multilevel modeling (Seltzer, 2004). The above discussion of regression modeling, however, indicates that regression-based treatment effect estimates can be sensitive to the parametric modeling assumptions, particularly if pre-treatment differences between treatment and control groups are large (Schafer & Kang, 2008). Preprocessing the data via matching can reduce the dependence on modeling assumptions when estimating treatment effects through a regression-based approach (Ho et al., 2007; Schafer & Kang, 2008). By taking a dual modeling approach in the analysis stage—where multilevel modeling is conducted with the matching-based preprocessed data—the proposed method seeks to utilize the advantages of multilevel modeling, while limiting dependence on modeling assumptions. Multilevel modeling can also be employed to examine factors associated with site-level treatment effect heterogeneity from a meta-analytic perspective (Seltzer, 2004), an option that the proposed method allows.

The literature in both the fields of causal inference and educational research only contain a handful of studies that incorporate both matching-based and regression-based methods to estimate causal effects in a multisite setting. The proposed method will add to this small, but growing, body of research. The primary contribution of the proposed method is that it extends the well-known concepts of matching and regression to a multisite setting where treatment assignment selectivity and heterogeneity complicate both regression-based and matching-based approaches. One of the few studies that examined the use of both multilevel propensity score models and multilevel models for treatment effect estimation (Su & Cortina, 2009) used a

within-site matching approach but suggested the use of a two-stage matching process to overcome complications with unmatched units.

The proposed method takes up this suggestion by extending the Stuart and Rubin (2008) two-stage matching approach to the multisite setting. This extension can be useful for educational researchers seeking to study programs or policies where treatment assignment can differ across sites and selection into different treatments is relatively selective. A prime example is the study of course-taking placements associated with curriculum differentiation, ability grouping, and tracking. In the following chapter I describe the proposed method in detail and lay out the process by which I demonstrate the method through an empirical illustration and evaluate the method through a series of simulation studies.

Chapter 3

The Two-Stage Matching Method and Study Design

The proposed two-stage matching method seeks to facilitate causal effect estimation in research settings complicated by the following factors: (1) random assignment is not practical or feasible; (2) assignment to the treatment condition is highly selective; (3) the assignment mechanism can vary across sites; and (4) the treatment effect can vary across units and sites. The first factor implies that a naïve treatment effect estimate will be a biased estimate of the true causal effect. The second factor implies that standard methods for selection bias adjustment may be ineffective and/or inefficient. For example, regression-based approaches may be highly dependent on parametric assumptions and extrapolation, and matching-based approaches may be limited by a lack of covariate overlap and common support. The third factor implies that any selection bias adjustment method based on the overall study sample or based on specific sites may not effectively correct for selection bias within any particular site. Lastly, the fourth factor implies that an estimate of the overall average causal effect may be less informative than an analysis of the heterogeneity in the causal effect across units and sites.

Settings where such factors are present can arise in educational research on policies or programs implemented across schools, where the policies/programs target a select population and implementation can vary across schools. A salient example that I use throughout the study is the use of differential course placement for students in the same grade-level, specifically placing some 8th graders in a pre-algebra course and others in a formal algebra course. Other examples could include school-based dropout prevention programs, tutoring or support services programs, and disciplinary correction programs.

In this chapter, I begin by describing the proposed method. I then outline the research questions for the study. Next, I describe the two techniques I employed to answer the research questions: an empirical illustration and simulation study.

3.1. The proposed two-stage matching method

To estimate causal effects under the research conditions described above, I propose using a method that addresses the complications in three phases. The first phase is the design phase, in which one uses a two-stage matching strategy to construct treatment and control groups that are well balanced along both unit- and site-level key pretreatment covariates. In the second phase, each non-local control unit's outcome is adjusted to account for potential site-level bias arising from the need to find matches outside the local site. The third phase is the analysis phase, in which one estimates average causal effects for the treated units and investigates heterogeneity in causal effects through multilevel modeling. The specific steps of each phase are a multisite extension of a single-site method described in Stuart and Rubin (2008). The first part of this section outlines the Stuart and Rubin method. I then elaborate on each phase of the proposed method.

3.1.1. The Stuart and Rubin (2008) multiple control group matching strategy

The Stuart and Rubin (2008) multiple control group matching strategy (henceforth the S&R method) was developed for a setting where a treatment is implemented within a single site but one can draw control units from both the local treatment site (control group 1) and a non-local non-treatment site (control group 2). More specifically, the S&R method is implemented through the following steps:

1. Match the treatment units to control units within the treatment site (control group 1), keeping only "good" matches based on a propensity score caliper range.
2. For matched control units from control group 1, find matches for them among the control units in control group 2.
3. For treatment units without a good match from step 1, find a match for them among the control units in control group 2.
4. Estimate the bias (υ) between the two control groups using a model estimated from the matched control group 1 and control group 2 units found in step 2, using, for example, a linear model for the observed control group outcome:

$$Y(0)_i = \beta_0 + \mathbf{\beta_1 X}_i + \upsilon C_i + e_i, \qquad e_i \sim N(0, \sigma^2),$$

where **X** is a $n \times p$ matrix of unit-level covariates and C is a dichotomous variable indicating whether the unit is in control group 1 or control group 2.

5. In preparation for the imputation of the counterfactual potential outcome, $Y(0)$, for control group 2 units draw:

$$s^2 \sim Inv - \chi^2(n-(p+2), \hat{\sigma}^2)$$

$$u \mid s^2 \sim N(\hat{\upsilon}, (\mathbf{X'X})^{-1} s^2),$$

where s^2 is the sampled error standard deviation and u is the sampled bias between the two control groups.

6. For each matched control unit from control group 2, adjust the observed outcome, $Y(0)$, by the estimated difference, u, between control group 1 and control group 2, so that $\hat{Y}(0)_i = Y(0)_i - u$ for all matched units in control group 2 and $\hat{Y}(0)_i = Y(0)_i$ for all matched units in control group 1.

7. Create a data set containing the observed outcome for all the treatment units, $Y(1)$, and the adjusted outcome for their matched control unit, $\hat{Y}(0)$.
8. Repeat steps 5 through 7 multiple times (i.e., create multiple data sets) to represent the uncertainty in the estimation of $\hat{\upsilon}$.

The eight steps outlined by S&R span the three phases of the proposed method. Steps 1 through 3 are in the design phase, where a two-stage matching approach is used to construct a single control group from a potential pool of two control groups (one local and one not local). Priority is given to focal matches in the local control group, but in the absence of focal local matches, focal matches are selected from the non-local control group. Steps 4 through 8 are in the adjustment phase, where the observed outcome for control group matches from the non-local group are adjusted to account for outcome differences between local and non-local sites. One can think of this adjustment as another counterfactual outcome for the non-local control group units: what would the unit's outcome under the control condition have been if the unit had been in the local site instead of the non-local site? This adjusted outcome is multiply imputed (step 8) to reflect the uncertainty we have in this counterfactual value. S&R discuss the analysis phase in steps 7 and 8, where one can estimate the average treatment effect based on the matched units' $Y(1)$ and $\hat{Y}(0)$ values (e.g., difference in means or a regression-adjusted difference) and the results based on the multiple data sets can be combined using standard multiple imputation combination rules (Little & Rubin, 2002).

3.1.2. The design phase: a two-stage matching strategy

For the research setting in question, the primary obstacle to overcome in the design phase is the highly selective nature of treatment assignment. Therefore, the primary objective in the

design phase is to construct a control group as similar as possible to the treatment group based on the important confounding pretreatment factors. In other words, one wants a focal local control group (Shadish & Cook, 2009) that will help rule out internal validity threats and facilitate site-specific average causal effects. To construct site-specific focal local control groups, I propose adapting the S&R method to a multisite setting. Employing this modified matching strategy allows one to preprocess (Ho et al., 2007) the data in the design phase to reduce dependence on modeling assumptions in the latter phases of the proposed method.

Two steps are required before one can execute the two-stage matching strategy in the design phase. The first step is to determine which pretreatment factors are important determinants of both treatment assignment and the outcome(s) of interest (Cochran, 1965) and, based on those factors, define a distance measure to use for matching (Stuart, 2010). A common distance measure is an estimated propensity score and, as discussed in section 2.3.5, propensity score estimation in a multisite setting is complicated by potential heterogeneity in the assignment mechanism across sites. Exploratory analysis of important pretreatment factors should include an investigation into the relative importance of site-level factors for both treatment assignment and the outcome(s). The extent to which site-level factors are important confounders relative to unit-level factors will help determine the extent to which local versus focal matches should be prioritized. Given important site-level heterogeneity, one could capture the heterogeneity in the propensity score model by using a separate single-level logistic regression model for each site if within-site sample sizes allow for reliable parameter estimates. However, this option many not be feasible, or efficient, for many situations in which one wishes to include a large number of covariates in the propensity score model and within-site sample sizes are limited. Another option is to use a multilevel logistic regression model to estimate each unit's propensity score, where

intercepts and slopes in the model are allowed to vary across sites (see Equation 2.8) and information from all sites is used to estimate site-specific estimates.

The second step is to identify site clusters from which one can draw acceptable between-site matches. The objective is to define groups of sites that increase the plausibility of the exchangeability assumption for control units in different sites. For example, for a study where students are nested within schools, schools could be clustered based on their geographic location, average pretreatment achievement levels, student demographics, or a combination of factors. By identifying site clusters and restricting between-site matching to other sites within the same cluster, one can limit the introduction of site-level bias into the two-stage matching process.

Once a propensity score is estimated for each unit and site clusters are identified, one can execute the two-stage matching process iteratively. The first stage is to conduct within-site matching (analogous to the S&R step 1). In this stage, treatment units within site *j* are matched to control units within site *j*, and only matches that fall within an acceptable propensity score caliper range are retained. A standard caliper range is 0.25 of a standard deviation (Rosenbaum & Rubin, 1985), but one can set the caliper range based on prior beliefs about the relative importance of focal and local factors. A wider caliper range will result in the retention of more within-site matches, but the treatment and control groups will not be as similar along the unit-level confounders. In other words, a wider caliper range sacrifices focal matches for local matches. Conversely, a narrower caliper range will result in within-site matches that are well balanced along the unit-level confounders but one may need to rely more on between-site matches (i.e., focal matches are prioritized over local matches). Matching can be conducted with or without replacement, but the use of replacement can complicate the adjustment and analysis

phases. The set of retained within-site matched units for site *j* is referred to as $M1_j$ (see top panel of Figure 3.1).

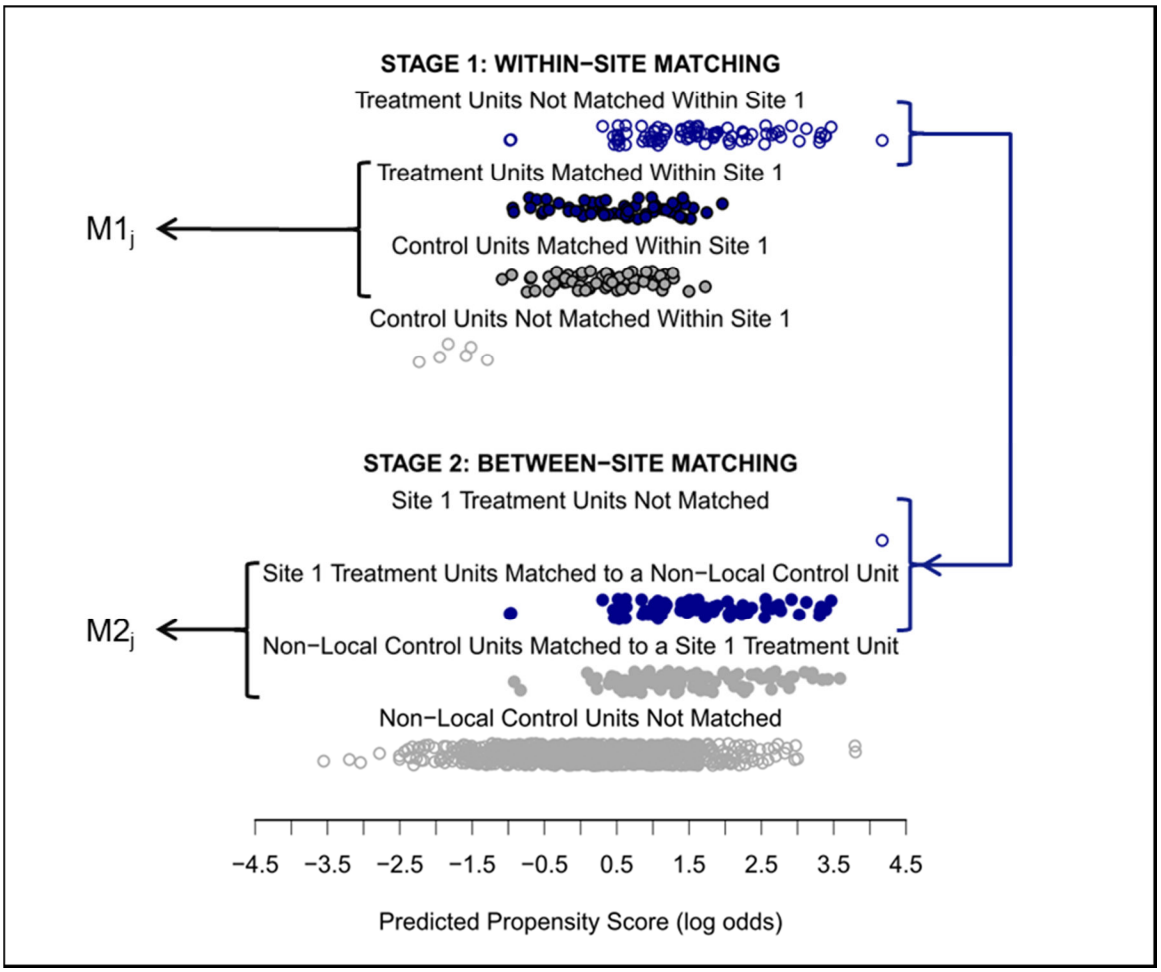

Figure 3.1. Jitter plots illustrating the two-stage matching strategy for one hypothetical site. Results from within-site matching (top panel) produce the $M1_j$ data set. Results from between-site matching (bottom panel) produce the $M2_j$ data set.

The second stage in the matching process is to find non-local matches for treatment units who were not retained in the within-site matching stage (analogous to the S&R step 3). Treatment units in site *j* who are not part of $M1_j$ are matched to control units who are not in site *j*

but are part of the same site cluster as site *j* (denoted as j'). However, if the propensity score model parameters vary across sites (e.g., if using a RIS multilevel logistic regression model), conducting between-site matching based on the predicted propensity score will not necessarily result in matching units with similar pretreatment covariate values. To overcome this potential mismatch problem, the predicted propensity score for control units who are not in site *j* can be re-estimated based on the propensity score model parameters specific to site *j*. For example, if each unit's propensity score was estimated from the multilevel logistic model represented in Equation 2.8, predicted propensity scores for units in site *j* would be based on the following:

$$\ln\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = \hat{\gamma}_{00} + \hat{\gamma}_{01}S_j^{gd} + \hat{u}_{0j} + (\hat{\gamma}_{10} + \hat{\gamma}_{11}S_j^{gd} + \hat{u}_{1j})(X_{ij}^{gp}),$$

with the predicted propensity scores for units in another site (j') within the same site cluster equal to

$$\ln\left(\frac{\hat{p}_{ij'}}{1-\hat{p}_{ij'}}\right) = \hat{\gamma}_{00} + \hat{\gamma}_{01}S_{j'}^{gd} + \hat{u}_{0j'} + (\hat{\gamma}_{10} + \hat{\gamma}_{11}S_{j'}^{gd} + \hat{u}_{1j'})(X_{ij'}^{gp}).$$

So the re-estimated propensity score for a control unit in site j' based on the site-specific parameter estimates for site *j* differs from the original predicted propensity score by the following amount (based on the log-odds transformation):

$$(\hat{u}_{0j} - \hat{u}_{0j'}) + \hat{\gamma}_{01}(S_j^{gd} - S_{j'}^{gd}) + (\hat{u}_{1j} - \hat{u}_{1j'} + \hat{\gamma}_{11}(S_j^{gd} - S_{j'}^{gd}))(X_{ij'}^{gp}).$$

If site clusters are constructed in a way that minimizes between-site differences in *S*, then adjustments involving ($S_j^{gd} - S_{j'}^{gd}$) will be minor. Additionally, note that the unit-level covariate, *X*, should be re-centered on the mean for site *j* instead of the mean for site j'.

Between-site matching can then proceed with the propensity scores re-estimated based on the target treatment site parameters. As with the within-site matching, only matches that fall

within an acceptable caliper range are retained. The set of retained between-site matched units for site *j* is referred to as $M2_j$ (see bottom panel of Figure 3.1). Ideally, the number of matched treatment units within the combined $M1_j$ and $M2_j$ data sets should equal the total number of treatment units in site *j*, or come close to the total number if some treatment units fall outside the common support range for all control units in the site cluster. For example, in the hypothetical match depicted in Figure 3.1, one treatment unit remains unmatched after the two-stage matching process. The proportion of treatment units in $M2_j$ provides a sense of how sensitive the resulting treatment effect estimates could be to the assumption of between-site exchangeability, with a high proportion of $M2_j$ treatment units meaning more dependence on between-site matches.

After implementing the two-stage matching process for site *j*, the process is repeated for all remaining sites (i.e., sites *j*+1 through *J*). This iterative matching process conceptually parallels a block randomized design, where treatment and control groups are created by random assignment implemented independently within each site. The resulting $M1_{j=1, 2, \ldots, J}$ and $M2_{j=1, 2, \ldots, J}$ data sets can be combined into a single matched data set which I will refer to as M. This data set contains all the matched treatment and control units from all sites.

As with other matching strategies, the quality of the resulting matched data set should be assessed based on whether covariate balance between the treatment and control groups improved after matching. The absolute standardized bias (*ASB*) and variance ratio (*VR*) for the estimated propensity score and each key covariate are useful summary statistics of covariate balance (Rubin, 2001):

$$ASB = \frac{\left|\bar{X}^t - \bar{X}^c\right|}{(\sigma^t + \sigma^c)/2} \tag{3.1a}$$

$$VR = \frac{\sigma^{2t}}{\sigma^{2c}} \tag{3.1b}$$

The objective in this phase is to not just create groups that are balanced on average, but groups that are balanced within each site. As a result, one should assess covariate balance for the whole sample and within each site. Remaining imbalance for specific sites could result in biased estimates of treatment effect heterogeneity. If the matching process did not result in acceptable improvements in covariate balance, the process should be repeated with adjustments made to the propensity score model specifications and/or the matching parameters (e.g., the caliper range or using replacement).

It is important to note that when cycling through the matching process for each site, the pool of control units can be replenished to ensure there are enough control units available to match to the treatment units in site *j*. If one conducts matching without replacement during the matching stage for a given site but replaces control units at the start of the matching stage for the next site, control units will be unique within any given $M1_j + M2_j$ data set but may appear multiple times across the matched data sets. The extent to which matched control units are duplicated across sites should be monitored because it could complicate subsequent phases of the method. If a small handful of control units are repeatedly matched to treatment units in different sites, it could be a sign that the overlap between treatment and control groups is particularly poor. When overlap is poor, it may be desirable to restrict the matching, and subsequent adjustment and analysis phases, to a subset of treatment units that fall within the range of control group common support. Restricting the treatment group sample in this way will redefine the population for which one can generalize the effect estimates, but reduces dependence on parametric assumptions and extrapolation in both the adjustment and analysis phases.

3.1.3. The adjustment phase: imputing counterfactual potential outcomes

The primary objective in the adjustment phase is to address the following counterfactual question for non-local control group matches: what would the observed outcome for control units in $M2_j$ have been if those units had been in site j instead of site j' . In using the two-stage matching strategy, matching some treatment units to control units in a non-local site may introduce bias in the treatment effect estimates if observed outcomes for control units differ across sites based on unobserved site-level factors. We can try to adjust for this bias prior to the analysis phase, however, by estimating site-level differences, or site effects, and extracting those differences from each control unit's observed outcome. Step 4 in the S&R method involves estimating the adjusted average difference between the local and non-local sites with a linear regression model based on an additional match among local and non-local control units.

This S&R step is complicated in the multisite setting in two particular ways. First, constructing a matched data set of control units based on the S&R method can, in the multisite setting, result in very few control units from specific sites. For example, if only 10 of 100 treatment units in a given site are part of the within-site match, the site effect estimates for that site will only be based on data from ten control units. To avoid this problem, site-effect estimation in the proposed two-stage matching process is based on all control units with a predicted propensity score within the same range as the matched control units. This strikes a balance between estimating school effects based on control units similar to the matched control units and maintaining a suitable sample size for precise site-specific model estimates.

Second, unlike the S&R method, the researcher needs to estimate unique site effects for each site, not just a single difference between a treatment site and non-local sites. For the multisite research setting, one can estimate each site's adjusted average effect with a hierarchical

model using the following random-intercept-and-slope hierarchical linear model for a continuous outcome:

$$\begin{aligned} &\text{Level 1: } Y(0)_{ij} = \beta_{0j} + \beta_{1j} X_{ij}^{c} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \\ &\text{Level 2: } \beta_{0j} = \gamma_{00} + \boldsymbol{\gamma_{01}} \mathbf{S}_j + u_{0j}, \quad u_{0j} \sim N(0, \tau_0), \\ &\qquad\qquad \beta_{1j} = \gamma_{10} + \boldsymbol{\gamma_{11}} \mathbf{S}_j + u_{1j}, \quad u_{1j} \sim N(0, \tau_1), \end{aligned} \tag{3.2}$$

where X_{ij}^{c} is a key pretreatment covariate, such as a pretest, for control unit *i* in site *j* centered around the site cluster mean, and $\mathbf{S}_j$ is a design matrix indicating the site cluster (defined in the matching stage) for site *j*. β_{0j} is the expected outcome for a control unit in site *j* who has a covariate value at the site cluster mean, and resides within a given site cluster. Similarly, β_{1j} is the expected linear relationship between the covariate and the outcome at site *j*. Site-level differences in expected outcomes for the average control group unit are captured by u_{0j}, and site-level differences in the expected relationship between the level-1 covariate and outcome measure are captured by u_{1j}. Estimates of these two random effects are empirical Bayes estimates of site effects, conditional on the site's cluster mean. The random effects are assumed to have a multivariate normal distribution with means zero, variance τ_0 and τ_1 respectively, and covariance captured by τ_{10} (not shown in Equation 3.2). This model is similar to what Raudenbush and Willms (1995) refer to as a nonuniform effects model for estimating school effects and what Reardon and Raudenbush (2009) refer to as a heterogeneous school effects model.

Based on the empirical Bayes model estimates, $u_{0j}^{*} + u_{1j}^{*} X_{ij}^{c}$ represents the expected effect of site *j* relative to other sites for a control unit with a given covariate value within the same site

cluster. Thus, the expected difference between unit *i*'s outcome if the unit had resided in site *j* instead of j' can be represented by the following:

$$(u^*_{0j} - u^*_{0j'}) + (u^*_{1j} - u^*_{1j'})X^c_{ij'}.$$

For each non-local matched control unit, we can use the above expected difference from the estimated model to adjust the control unit's outcome for the counterfactual condition of residing in the local site (*j*) instead of the non-local site (j'):

$$E[Y(0)_{ij} \mid Y(0)_{ij'}, X^c_{ij'}, \mathbf{S}_j, \hat{\boldsymbol{\theta}}] = \tilde{Y}(0)_{ij} = Y(0)_{ij'} + (u^*_{0j} - u^*_{0j'}) + (u^*_{1j} - u^*_{1j'})X^c_{ij'}, \quad (3.3)$$

where $\hat{\boldsymbol{\theta}}$ is the vector of model parameters estimated from the above hierarchical model. One can read the left-hand side expectation in Equation 3.3 as the expected outcome value under the control condition for unit *i* if unit *i* had resided in site *j*, given: the observed outcome value under control condition for unit *i* in site j' , the covariate value for unit *i* in site j' , the site cluster sites *j* and j' reside, and the estimated model parameters. In other words, $\tilde{Y}(0)_{ij}$ is the expected local site counterfactual outcome for a non-local control unit.

The rationale and logic for the above adjustment can be illustrated with a simple example where some treatment group students in school A are matched to control group students in school A (the M1 match in Figure 3.1), and other treatment group students in school A are matched to control group students in schools B and C (the M2 match in Figure 3.1). All three schools are in the same site cluster. Using all control units that fall within the propensity score range of the matched control units, site effects are estimated based on Equation 3.2. The estimated site effects for six hypothetical control students matched to school A treatment students are displayed in Table 3.1. For this example, assume the grand-mean outcome value for the site cluster is 50 (i.e., $\gamma_{00} + \gamma_{01} = 50$) and the site cluster grand-mean slope for the covariate is

10 (i.e., $\gamma_{10} + \gamma_{11} = 10$). School A has an estimated adjusted mean 10 points below the cluster mean and a slope equal to the cluster mean slope. School B, however, has an estimated adjusted mean 15 points above the cluster mean and a weaker relationship between the covariate and outcome, while school C has an estimated adjusted mean equal to the cluster mean and a stronger relationship between the covariate and outcome. Given these estimates, the two non-local control units with a covariate value of 1 (students 3 and 5) have their potential outcome adjusted downward from 70 to 50, which matches the observed outcome for the school A control student with the same covariate value (student 1). For the two non-local control units with a covariate value of -1 (students 4 and 6), the potential outcome for student 4 is adjusted to 30, while student 6's adjustment is zero. The adjustment for the six hypothetical control group students can also be illustrated graphically. Figure 3.2 shows the empirical Bayes fitted regression lines for each school, and how the adjustments "move" students from the fitted line for their school to the fitted line for School A. In doing so, the observed outcome for non-local control units is adjusted to reflect the counterfactual outcome if the control units had been in School A.

Table 3.1. Hypothetical example of counterfactual potential outcome adjustment for six control group students matched to treatment group students in School A.

Student ID	School ID	Observed Outcome ($Y(0)$)	X^c	u^*_{0A}	$u^*_{0j'}$	u^*_{1A}	$u^*_{1j'}$	Adjustment	Adjusted Outcome ($\tilde{Y}(0)$)
1	A	50	1	-10	-10	0	0	0	50
2	A	30	-1	-10	-10	0	0	0	30
3	B	70	1	-10	15	0	-5	-20	50
4	B	60	-1	-10	15	0	-5	-30	30
5	C	70	1	-10	0	0	10	-20	50
6	C	30	-1	-10	0	0	10	0	30

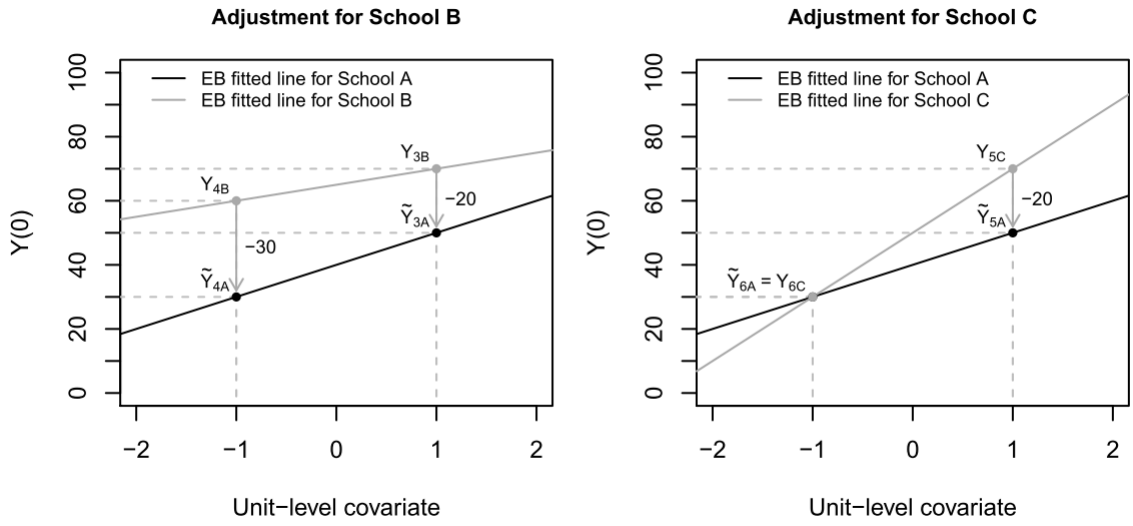

Figure 3.2. Hypothetical example of counterfactual potential outcome adjustment for non-local control unit matched to treatment units in School A.

Given that the difference between $\tilde{Y}(0)_{ij}$ and $Y(0)_{ij}$ depends on the model-estimated site-level random effects, one may find it desirable to incorporate uncertainty in the random effect estimates into the analysis. As in step 5 of the S&R method, this uncertainty can be incorporated by imputing $\tilde{Y}(0)_{ij}$ multiple times based on random draws from a model-estimated distribution of random effect estimates. To do this based on the multilevel model presented in Equation 3.2, one can first draw a sample random effects variance-covariance matrix, $\mathbf{T}^m$, from the following inverse-Wishart distribution with given degrees of freedom and scale matrix:

$$\mathbf{T}^m \sim Inv-Wishart(df, df(\hat{\mathbf{T}})), \tag{3.4a}$$

where $\hat{\mathbf{T}}$ is the 2×2 model-estimated variance-covariance tau matrix and $\text{df} = J - (q+2)$ when J is the total number of sites and q is the number of level-2 covariates. Then, given the sampled

tau matrix, one can sample random effect estimates from the following multivariate normal distribution:

$$\mathbf{U}_j^m \sim MVN(\mathbf{U}_j^*, \mathbf{V}_j^*), \tag{3.4b}$$

where $\mathbf{U}_j^*$ is a 2 × 1 vector of the model-estimated empirical Bayes random effects for site *j* and $\mathbf{V}_j^*$ is a 2 × 2 conditional variance-covariance matrix of the random effects. Following Chapter 3 in Raudenbush and Bryk (2002), the $\mathbf{V}_j^*$ matrix is the posterior variance-covariance matrix for the empirical Bayes random effects given the data and variance components, and can be estimated with the following formula:

$$\mathbf{V}_j^* = (\hat{\mathbf{V}}_j^{-1} + \mathbf{T}^{-1})^{-1} + (\mathbf{I} - \mathbf{\Lambda}_j)(\mathbf{S}\hat{\mathbf{V}}_{\hat{\gamma}}\mathbf{S}')(\mathbf{I} - \mathbf{\Lambda}_j)', \tag{3.4c}$$

where **T** is now the sampled random effects variance-covariance matrix ($\mathbf{T}^m$). $\hat{\mathbf{V}}_j$ is the 2 × 2 ordinary least squares estimated random effects variance-covariance matrix for site *j*,

$$\hat{\mathbf{V}}_j = \hat{\sigma}^2(\mathbf{X}_j'\mathbf{X}_j)^{-1},$$

where **X** is a n_j × 2 matrix with the ones in the first column for the intercept and each unit's cluster mean centered propensity score in the second column. $\mathbf{\Lambda}_j$ is the 2 × 2 multivariate reliability matrix,

$$\mathbf{\Lambda}_j = \mathbf{T}(\mathbf{T} + \hat{\mathbf{V}}_j)^{-1}.$$

$\hat{\mathbf{V}}_{\hat{\gamma}}$ is the estimated sampling variance for the level-2 predicted values,

$$\hat{\mathbf{V}}_{\hat{\gamma}} = \left(\sum \mathbf{S}_j'(\mathbf{T} + \hat{\mathbf{V}}_j)^{-1}\mathbf{S}_j\right)^{-1},$$

where **S** is a 2 × 2(q+1) design matrix for the level-2 intercept and covariates.

With *M* draws from the conditional posterior multivariate distribution for the empirical Bayes random effects, the adjustment phase concludes with the researcher imputing *M*

counterfactual values for each of the non-local control units based on the adjustment in Equation 3.3. This produces *M* data files to use for estimation of the average treatment effect for the treated. Alternative methods for multiply imputing the adjusted outcome values, particularly those using Markov chain Monte Carlo, are possible but not addressed in this study. Regardless of the imputation method, estimation of the treatment effect in the analysis stage can be conducted separately for each of the multiply imputed data files and the results can be combined using standard multiple imputation combination rules (Little & Rubin, 2002), which will be discussed in more detail in the following section.

3.1.4. The analysis phase: estimating average treatment effects and investigating effect heterogeneity

In the analysis phase, one utilizes the *m* data sets from the adjustment phase to estimate average treatment effects. Using these data, one can make three types of inferences. The first is an inference about the overall average treatment effect for the treated units (ATT), or what Gitelman (2005) referred to as the group-allocation, multilevel average (GAMA). The second inference is about the degree of ATT site-level variance. The third type of inference is about what factors are associated with the site-level variance.

With the matched data, one can use an unconditional random-intercept-and-slope hierarchical model to address inferences about the GAMA and variation in the site-level ATT, where the dependent variable is the observed outcome for each treatment unit and the adjusted outcome for each control unit based on imputation *m*:

$$
\begin{aligned}
&\text{Level 1: } Y_{ij} = D_{ij}Y(1)_{ij} + (1 - D_{ij})\tilde{Y}(0)^{m}_{ij} = \beta_{0j} + \beta_{1j}D_{ij} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \\
&\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_0), \\
&\qquad\qquad \beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_1).
\end{aligned}
\tag{3.5}
$$

In Equation 3.5, γ_{10} represents the estimated GAMA, or the grand-mean ATT. Each site-level ATT is captured by β_{1j} and the degree to which site-level ATTs vary around the grand-mean ATT is captured by τ_1.

With an estimate of the GAMA and ATT site-level variance, one can examine factors associated with treatment effect heterogeneity. This can be accomplished by adding site-level mediator and/or moderator variables to the above unconditional hierarchical model. For example, to test whether the average effect of taking an above-grade-level course differs across sites that encourage (*E*) versus discourage students to take an above-grade-level course, one can use the following model:

$$
\begin{aligned}
&\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}D_{ij} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \\
&\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}E_j + u_{0j}, \quad u_{0j} \sim N(0, \tau_0), \\
&\qquad\qquad \beta_{1j} = \gamma_{10} + \gamma_{11}E_j + u_{1j}, \quad u_{1j} \sim N(0, \tau_1).
\end{aligned}
\tag{3.6}
$$

where γ_{10} now represents the GAMA for sites that discourage above-grade-level course-taking (i.e., *E*=0) and γ_{11} is the degree to which the GAMA differs, on average, for sites that encourage above-grade-level course-taking. The amount of site-level effect variance explained by the encouragement measure can be assessed by comparing the estimate of τ_1 in the above model to the estimate from the unconditional model.

The above discussion of the analysis phase describes how one can estimate average treatment effects and effect heterogeneity for a single imputed data set. However, when working with multiple data files that reflect multiple imputations of the counterfactual outcomes,

inferences should be based on results combined across the *M* data files. So estimation of the parameters of interest (e.g., the ATT) can be conducted separately for each data file and the results can be combined using standard multiple imputation combination rules (Little & Rubin, 2002). The overall point estimate for a given parameter, θ, is simply the mean value across all the imputed data files:

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^{m},$$

where *M* is the total number of imputed data files and *m* is an index for the *m*=1,…, *M* data files. While the uncertainty in $\bar{\theta}$ consists of two parts: an average of the within-imputation variance,

$$\bar{W} = \frac{1}{M}\sum_{m=1}^{M}W^{m},$$

and the between-imputation variance,

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\theta^{m} - \bar{\theta})^{2}.$$

The within- and between-variance are combined to get total uncertainty in $\bar{\theta}$ based on the following formula:

$$T = \bar{W} + \frac{1+M}{M}B.$$

Additionally, interval estimates and significance tests can be based on a *t* distribution,

$$\frac{(\theta - \bar{\theta})}{\sqrt{T}} \sim t_{v},$$

with degrees of freedom equal to:

$$(M-1)\left(\frac{1}{M+1}\frac{\bar{W}}{B}\right)^{2}.$$

3.1.5. Key assumptions in the proposed method

As with all causal inference methods, applying the proposed method will only result in valid, unbiased, treatment effect estimates if specific key assumptions hold. The assumptions fall into three categories: strongly ignorable treatment assignment, SUTVA, and modeling assumptions. These key assumptions are outlined below.

Strongly ignorable treatment assignment. As discussed in the previous chapter, unbiased estimation of treatment effects in a non-experimental study requires that the potential outcomes are independent of treatment assignment given the observed covariates: $[Y(1), Y(0)] \perp D \mid X$. This is the same as assuming selection on the observables or that the potential outcomes are missing at random (MAR). While the validity of this assumption cannot be fully tested, equating treatment and control units along a dimension of pre-treatment factors known to be important confounders increases the plausibility of this assumption. Furthermore, matching units within sites ensures that site-level factors will not invalidate the assumption. However, for the proposed between-site matching step to be effective, one must assume that control units within the site clusters are exchangeable after adjusting for the observed confounders and the estimated site-level effects. Referring back to Equation 3.3, this means $E[Y(0)_{ij} - \tilde{Y}(0)_{ij}] = 0$ for a non-local control unit. By multiply imputing $\tilde{Y}(0)_{ij}$ we are able to incorporate some uncertainty about the validity of this assumption into the analysis.

SUTVA. By focusing on Gitelman's (2005) GAMA and the site-level ATT, the proposed methodology relaxes SUTVA as it pertains to a single-site study to allow for heterogeneity in treatment enactment and potential outcomes across sites. Valid estimation of the site-level ATT still requires application of SUTVA within each site and an assumption of no interference between sites. It also requires assumptions about how units are allocated to sites. Namely, site

membership is assumed to be invariant (i.e., remain stable and intact) during the treatment period.

Modeling assumptions. As with any statistical analysis, validity of the results depend on whether certain modeling assumptions hold. The main modeling assumptions are assumptions about a model's functional form, that the error terms are independent and identically distributed (i.i.d.), and that distributional assumptions about the observed and latent variables hold. For the proposed method, these assumptions are part of the propensity score model in the design phase, the site-effects model in the adjustment phase, and the effect estimation models in the analysis phase. The proposed method is designed to minimize bias due to misspecification of the functional form by using the matched treatment and control units, where covariate overlap and common support are strong, in the adjustment and analysis phases.

In addition to the above assumptions, it is important to recognize that for simplicity, the proposed method does not address a number of non-trivial issues that can result in biased treatment effect estimates. For example, units are assumed to fully comply with the treatment in question and any attrition from the study is assumed to be independent of treatment assignment. Additionally, the proposed method proceeds as if the observed data for the multisite study do not include missing values or measurement error. Standard approaches to remedy these common problems could be incorporated into the proposed method, however.

3.2. Research questions

To test the utility and feasibility of the proposed method, this study addressed the following research questions:

1. How do different specifications in the design phase of the proposed method influence covariate balance?
2. How do different specifications in the design phase of the proposed method influence inferences about treatment effects?
3. How do different specifications in the adjustment phase of the proposed method influence inferences about treatment effects?
4. How do treatment effect estimates from the proposed method compare to estimates obtained from more common matching-based and regression-based methods?

I employed two techniques to address the above research questions. First, I used a set of Monte Carlo simulation studies to compare the performance of the proposed method to different specifications and methods under conditions where treatment selection and treatment effects are known. Second, I applied the proposed method to an empirical data analysis to illustrate the proposed method and its results relative to commonly used methods. The simulation study and empirical illustration are described in the next sections.

3.3. Methods for simulation study

I used a set of Monte Carlo simulation studies to address research questions 1 through 4. The main purpose of the simulation studies was to investigate the method's sensitivity to different specifications within each of the three phases under different assignment mechanism conditions. If results are robust to certain specifications within a phase, then the proposed method can be implemented with the more parsimonious, or implementable, specifications. I treated the examination of specifications within each of the three phases as separate simulation

studies, where, for a given phase, the conditions in the other two phases were fixed. Each study was based on 100 Monte Carlo replications with five assignment mechanism conditions and additional conditions specific to each phase. The sample size for each replication was fixed to have 50 sites in a given data set ($J = 50$). To allow for unbalanced groups, the within-site sample size (n_j) for each site was based on a draw from the following normal distribution: $n_j \sim N(200,10)$. The proposed method is assumed, *a priori*, to be inappropriate for small sample size settings where few within-site matches are feasible (e.g., $n_j = 50$) and where between-site heterogeneity is difficult to estimate (e.g., $J = 20$). Future analyses will compare performance of the two-stage matching method under the fixed sample size to larger sample size conditions.

3.3.1. Measures of performance

To summarize performance of each simulation condition across the Monte Carlo replications, I used a series of common measures. For research question 1, the focus was on covariate balance. To assess balance, I looked at the average within-site *ASB* and *VR* (see Equations 3.1a and 3.1b, respectively) across the simulation replications. *ASB* and *VR* between-replication variance (i.e., the Monte Carlo standard deviation) was also monitored to gauge precision of the Monte Carlo estimates. With 100 replications, statistically significant differences in performance between conditions can be detected if the differences are approximately 0.20 of a standard deviation apart. In addition to overall balance on these two measures, of particular interest was how well these balance measures held within each site, which was assessed by the mean between-site variance in the *ASB* and *VR* across the simulation replications, as well as the mean maximum site-level ASB and VR across replications. For research questions 2 through 4, the focus was on treatment effect estimates, particularly the grand-mean ATT (GAMA) and site-

level ATT effect variance. To assess how well each condition recovers the true GAMA and site-level ATT variance, I looked at estimator bias and root mean squared error (RMSE). Bias was measured as the average difference between the estimated parameter value and the true parameter value across the simulation replications:

$$\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}-\theta), \tag{3.7a}$$

where R is the number of Monte Carlo replications, $\hat{\theta}$ is the estimated parameter value of interest for a given replication, and θ is the true parameter value for a given replication. Similarly, RMSE for a give parameter value is measured based on the following formula:

$$\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}-\theta)^2}. \tag{3.7b}$$

I also examined coverage, or how often a ± 2 × standard error interval around the estimated parameter value contained the true parameter value, across the replications.

3.3.2. Simulation conditions: treatment assignment mechanism

Five assignment mechanism conditions were selected to cover the different assignment mechanisms discussed in Chapter 2 (see Figures 2.1 and 2.2). The different assignment mechanisms allow one to compare the proposed method's performance under selection independence, selection on unit- and site-level observables, and selection on unit- and site-level unobservables. For all conditions, treatment assignment for a given unit was based on a draw from a binomial distribution, where each unit's probability of being assigned to the treatment condition ($\Pr(D_{ij}=1)=p_{ij}$) was determined by the following multilevel logistic model:

$$\text{Level 1: } \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{ij} + \beta_{2j}Z_{ij} + \beta_{3j}U_{ij}, \tag{3.8}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}S_j + \gamma_{02}V_j,$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}S_j + \gamma_{12}V_j,$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}S_j + \gamma_{22}V_j,$$
$$\beta_{3j} = \gamma_{30} + \gamma_{31}S_j + \gamma_{32}V_j,$$

where X and Z are observed unit-level covariates, U is an unobserved unit-level covariate, S is an observed site-level covariate, and V is an unobserved site-level covariate. The model's parameter values differed across the five assignment mechanism conditions:

- Random assignment;
- Selection on unit-level observables;
- Selection on unit- and site-level observables;
- Selection on unit-level observables and site-level observables and unobservables;
- Selection on unit- and site-level observables and unobservables.

The actual model parameter values used for each assignment mechanism condition are provided in Table 3.2.

The proposed method is hypothesized to perform as well as other standard effect estimation methods under the random assignment and level-1 observable conditions. Under the other conditions, where site-level factors influence the assignment mechanism, the proposed method is hypothesized to result in more valid inferences regarding the treatment effect. Given the potential for within-site matching and site-effect adjustment with multilevel models, the proposed method is hypothesized to perform best, relative to standard effect estimation methods, under conditions where the assignment mechanism depends on unit-and site-level observables

plus site-level unobservables. No method is hypothesized to perform very well under conditions that include unobservables at both the unit- and site-levels.

Table 3.2. Probability of treatment assignment model parameter values for different assignment mechanism conditions (see Equation 3.8 for model).

	Random Assignment	Unit Observable	Unit & Site Observable	Unit & Site Observable + Site Unobservable	Unit & Site Observable + Unit & Site Unobservable
Parameters for β_{0j}					
γ_{00}	-0.50	-0.50	-0.50	-0.50	-0.50
γ_{01}	0.00	0.00	0.40	0.40	0.40
γ_{02}	0.00	0.00	0.00	0.20	0.20
Parameters for β_{1j}					
γ_{10}	0.00	0.50	0.50	0. 50	0.50
γ_{11}	0.00	0.00	0.30	0.30	0.30
γ_{12}	0.00	0.00	0.00	0.25	0.25
Parameters for β_{2j}					
γ_{20}	0.00	0.40	0.40	0.40	0.40
γ_{21}	0.00	0.00	0.20	0.20	0.20
γ_{22}	0.00	0.00	0.00	0.20	0.20
Parameters for β_{3j}					
γ_{30}	0.00	0.00	0.00	0.00	0.50
γ_{31}	0.00	0.00	0.00	0.00	0.30
γ_{32}	0.00	0.00	0.00	0.00	0.25

3.3.3. Simulation conditions: data generation

Aside from the treatment assignment model, the data generation models were fixed for all simulation conditions. The data generation models were chosen to represent a setting where

potential outcomes are influenced by two observed factors and an unobserved factor at the unit-level and an observed and unobserved factor at the site-level. As a contextual reference, consider potential outcomes under a treatment or control condition ($Y(1)_{ij}$ and $Y(0)_{ij}$) on a mathematics standardized test score for students nested within schools, where the researcher has an observed measure of prior academic achievement for each student (X_{ij}), an observed measure of student socio-economic status (Z_{ij}), and an observed composite measure of each school's overall instructional resources (S_j). Also important in determining a student's test score outcome, however, are an unobserved student motivation factor (U_{ij}) and an unobserved school instructional quality factor (V_j). The data generation models were also chosen to represent a setting where unit-level factors are not evenly distributed across sites. In other words, school factors are correlated with student factors. Parameter values for the data generation and treatment assignment models were determined after exploratory testing of appropriate and plausible values.

To generate data that represents the above condition, the two site-level variables were drawn from normal distributions with zero means and standard deviations of one:

$$S_j \sim N(0,1),$$
$$V_j \sim N(0,1).$$

Then the three unit-level variables were drawn from the following normal distributions:

$$X_{ij} \sim N(\gamma_{x1}S_j + \gamma_{x2}V_j, \sigma_x^2), \text{ where } \gamma_{x1} = 0.20,\ \gamma_{x2} = 0.10,\ \sigma_x^2 = 0.70,$$
$$Z_{ij} \sim N(\gamma_{z1}S_j + \gamma_{z2}V_j, \sigma_z^2), \text{ where } \gamma_{z1} = 0.10,\ \gamma_{z2} = 0.20,\ \sigma_z^2 = 0.70,$$
$$U_{ij} \sim N(\gamma_{u1}S_j + \gamma_{u2}V_j, \sigma_u^2), \text{ where } \gamma_{u1} = 0.10,\ \gamma_{u2} = 0.20,\ \sigma_u^2 = 0.70.$$

With values for both site- and unit-level variables, the potential outcome under control condition was based on the following multilevel model:

$$
\begin{aligned}
Y(0)_{ij} &= \beta_{0j} + \beta_{1j}X_{ij} + \beta_{2j}X_{ij}^2 + \beta_{3j}X_{ij}^3 + \beta_{4j}Z_{ij} + \beta_{5j}U_{ij} + e_{ij}, \text{ where } e_{ij} \sim N(0, 0.25) \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}S_j + \gamma_{02}V_j, \text{ where } \gamma_{00} = 0.00, \gamma_{01} = 0.50, \gamma_{02} = 0.30, \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}S_j + \gamma_{12}V_j, \text{ where } \gamma_{10} = 0.75, \gamma_{11} = 0.30, \gamma_{12} = 0.20, \\
\beta_{2j} &= \gamma_{20}, \text{ where } \gamma_{20} = 0.05, \\
\beta_{3j} &= \gamma_{30}, \text{ where } \gamma_{30} = 0.05, \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}S_j + \gamma_{42}V_j, \text{ where } \gamma_{40} = 0.50, \gamma_{41} = 0.10, \gamma_{42} = 0.20, \\
\beta_{5j} &= \gamma_{50} + \gamma_{51}S_j + \gamma_{52}V_j, \text{ where } \gamma_{50} = 0.50, \gamma_{51} = 0.20, \gamma_{52} = 0.30.
\end{aligned}
$$

This model let the relationships between the unit-level covariates and the potential outcome differ across sites based on the site-level variables. Additionally, the model included a non-linear relationship between *X* and the potential outcome to allow for a condition that cannot be perfectly modeled by a simple linear regression. The non-linear component was fixed across sites. Lastly, the potential outcome under a treatment condition was based on the following model:

$$
\begin{aligned}
Y(1)_{ij} &= Y(0)_{ij} + \delta_j, \\
\delta_j &= \gamma_{00} + \gamma_{01}X_{ij} + \gamma_{02}S_j + \gamma_{03}V_j, \\
&\text{where } \gamma_{00} = 0, \gamma_{01} = 0.75, \gamma_{02} = 0.50, \gamma_{03} = 0.25.
\end{aligned}
\tag{3.9a}
$$

This model allowed for treatment effect heterogeneity across sites based on the two-site level variables (*S* and *V*) and across units within sites based on *X*. The observed outcome for a given unit (Y_{ij}) was determined by each unit's treatment assignment condition (D_{ij}):

$$
Y_{ij} = D_{ij}Y(1)_{ij} + (1 - D_{ij})Y(0)_{ij}, \tag{3.9b}
$$

where D_{ij} took on a value of one or zero depending on the assignment mechanism models described by Equation 3.10.

3.3.4. Design phase simulation study

To examine performance of the proposed method across different specifications in the design phase (research questions 1 and 2), the simulation study examined two decision points within the design phase. The main decision in the design phase is to preprocess the data with matching or proceed to the analysis phase with the full data sample (i.e., skip the design and adjustment phases). Given the decision to match, the simulation study examined two decision points in the matching process: propensity score model specification and matching method specification.

Three different propensity score model conditions were included in the simulation to examine how using a multilevel model to estimate the propensity score influences covariate balance and treatment effect estimates. The three propensity score model conditions were:

- A single-level logistic regression model with *X, Z* and *S* as predictors;
- A two-level random intercept (RI) logistic regression model with *X* and *Z* as level-1 predictors; and
- A two-level random intercept and slope (RIS) logistic regression model with *X* and *Z* as level-1 predictors and *S* as a level-2 predictor for both the intercept and slopes.

It was hypothesized that the RIS model would perform better than the other two models under the three assignment mechanism conditions that include site-level factors. The single-level and RI models, however, were hypothesized to perform better than the RIS model under the two assignment mechanism conditions that did not include site-level factors.

In all simulation conditions that include matching, I used a 1-to-1 caliper match with the caliper set at 0.25 standard deviations of the propensity score. While a wide variety of matching

method specifications are possible, the main specification of interest for the proposed method was whether a two-stage process improves covariate balance and effect estimation. Therefore, I examined three different ways to match units based on a given propensity score and the 1-to-1 caliper match process:

- Matching that ignored the nesting of units within sites and allowed units to be matching within or between sites (pooled matching);
- Matching that was restricted to within-site matches (i.e., exact matching on site); and
- The two-stage matching method that prioritized within-site matches then looked for between-site matches.

Crossed with the assignment mechanism conditions, the design phase simulation study included 45 conditions (40 matching conditions + 5 no match conditions). Since within-site matches should produce groups with both identical observed and unobserved site-level characteristics, it was hypothesized that the within-site matching method and the two-stage matching method would perform relatively better under the two assignment conditions that included the unobserved site-level factor. Additionally, it was hypothesized that the two-stage matching method would retain more treatment units than the within-site matching method, but would do so at the expense of non-exact matches on site-level factors. The same design phase simulation conditions were used to address research question 2. For each simulation condition, the GAMA and its associated site-level variance were estimated using a two-level unconditional regression model in the analysis phase. For the two-stage matching condition, five counterfactual imputations in the adjustment phase were used prior to the analysis phase. It was hypothesized

that the two-stage matching method would perform best under conditions where treatment assignment depends on both unit- and site-level covariates.

3.3.5. Adjustment phase simulation study

The objective in the adjustment phase is to impute counterfactual values for non-local site control group units to remove residual site effects. The main decision to make at this phase is how many imputations, if any, to estimate. To examine performance of the proposed method across different imputation specifications in the adjustment phase (research question 3), the simulation study had four simulation conditions for imputation of non-local control group counterfactuals:

$$m = 0, 1, 5, \text{ or } 10.$$

Crossed with the assignment mechanism condition, the imputation phase simulation study included 20 conditions. For each simulation condition, the GAMA and its associated site-level variance were estimated with the two-stage matching strategy in the design phase—using a two-level RIS propensity score model and 0.25 caliper—and a two-level unconditional regression model in the analysis phase. It was hypothesized that under conditions where treatment assignment depended on both unit- and site-level covariates, imputing the non-local control group counterfactuals would reduce bias. Additionally, multiply imputing the counterfactual values adds a layer of complexity in both the adjustment and analysis phase, namely drawing plausible school effect values and combining estimates across the imputations, so attention was given to the performance of the single imputation condition relative to the multiply imputed condition.

3.3.6. Analysis phase simulation study

Given matching in the design phase and imputation of control counterfactuals in the adjustment phase, specifications in the analysis phase are relatively straightforward. If the matching operates as theorized, average treatment effects can be estimated with a simple difference in group means. If matching does not result in balanced groups, however, different outcome model specifications may be combined with matching to adjust for residual covariate bias (i.e., dual modeling). Additionally, it is possible that a more straightforward, and traditional, regression-based approach could produce unbiased effect estimates without matching in the design phase. Thus, the analysis phase simulation study examined how treatment effect estimation based on the proposed two-stage matching method differed across outcome model specifications in the analysis phase and compared to estimation from more traditional regression-based effect estimation methods (research question 4).

In particular, I compared five different regression model specifications for effect estimation crossed with either the unmatched original data or preprocessed data using the two-stage matching method. The five regression model conditions are described in Table 3.3. Crossed with the assignment mechanism conditions, the analysis phase simulation study included 50 conditions. It was hypothesized that all analysis phase conditions would perform well under random assignment. Since matching should at least partially account for bias arising from unobserved between-site differences and misspecification of the functional form, I also hypothesized that conditions based on data preprocessed with the two-stage matching method would have less bias than the regression-only estimates under all other assignment mechanism conditions. In addition to the results from the analysis phase simulation study, results from both the design phase and imputation phase studies help address research question 4.

Table 3.3. Regression model conditions for treatment effect estimation in the analysis phase simulation study.

Condition	Model
1. Single-level model with treatment group indicator	Level 1: $Y_{ij} = \beta_0 + \beta_1 D_{ij} + e_{ij}$
2. Single-level model with treatment group indicator and controls for observed covariates	Level 1: $Y_{ij} = \beta_0 + \beta_1 D_{ij} + \beta_2 X_{ij}^{gd} + \beta_3 Z_{ij}^{gd} + \beta_4 S_j^{gd} + e_{ij}$
3. Two-level RIS model with treatment group indicator at level-1	Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + e_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$
4. Two-level RIS model with treatment group indicator and control for unit-level observed covariates	Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + \beta_{2j} X_{ij}^{gp} + \beta_3 Z_{ij}^{gp} + e_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$ $\beta_{2j} = \gamma_{20} + u_{2j}$ $\beta_{3j} = \gamma_{30} + u_{3j}$
5. Two-level RIS model with treatment group indicator and control for observed unit- and site-level covariates	Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + \beta_{2j} X_{ij}^{gp} + \beta_3 Z_{ij}^{gp} + e_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} S_j^{gd} + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11} S_j^{gd} + u_{1j}$ $\beta_{2j} = \gamma_{20} + \gamma_{21} S_j^{gd} + u_{2j}$ $\beta_{3j} = \gamma_{30} + \gamma_{31} S_j^{gd} + u_{3j}$

3.4. Methods for empirical illustration

The empirical illustration was designed to serve three purposes. First, findings from the illustration informed research question 4 by comparing results based on the proposed method to results from more traditional analytic methods. Second, the illustration tested the feasibility and utility of the proposed method for addressing real-world educational policy questions with non-experimental multisite data. Third, the illustration demonstrated how the proposed method facilitates exploration of treatment effect heterogeneity across units and sites. To accomplish

these three objectives, the empirical illustration utilized longitudinal student data from a large urban school district to examine the effect of mathematics course differentiation in 8th grade on high school mathematics achievement. This section describes the data and how I used these data to test the proposed method.

3.4.1. Data for empirical illustration

Since the proposed method seeks to facilitate causal effect estimation in research settings complicated by selective treatment assignment, where both the assignment mechanism and treatment effects may vary across sites, I identified an empirical illustration that met this condition. Specifically, the data cover 19,063 students, within 50 schools, who were 8th graders in a large urban school district during the 2003-04 school year. These students were assigned to one of two mathematics courses in 8th grade: (1) a course designed to cover a full year of Algebra 1 content material in 8th grade (treatment condition), or (2) a slower paced mathematics course designed to only cover the first half of Algebra 1 content material in 8th grade (control condition). For the slower paced course, the second half of Algebra 1 content is expected to be covered in 9th grade. Since a large portion of the slower paced course focused on pre-Algebra content, for simplicity, I refer to this control condition course as pre-Algebra and the treatment condition course as Algebra. Across this cohort, 6,330 students (33%) took the Algebra course and 12,733 students took the pre-Algebra course. It is important to note that the treatment in this illustration is defined as assignment to an Algebra course at the start of 8th grade instead of a pre-algebra course, which may differ from a treatment defined as receipt of a full year of algebra content versus a full year of pre-algebra content.

The total number of students and the proportion of students assigned to Algebra varied dramatically across schools within the district. To facilitate the feasibility of selecting within-school matches and estimating within-school effects, I restricted the cohort to schools that had at least 50 8th graders and between 10% and 90% of students in the treatment condition. Among these 50 schools, the total number of students in the cohort ranged from 53 to 973 students, while the percent of students assigned to Algebra ranged from 11% to 87% (see Figure 3.3 for how the schools are distributed across these two characteristics).

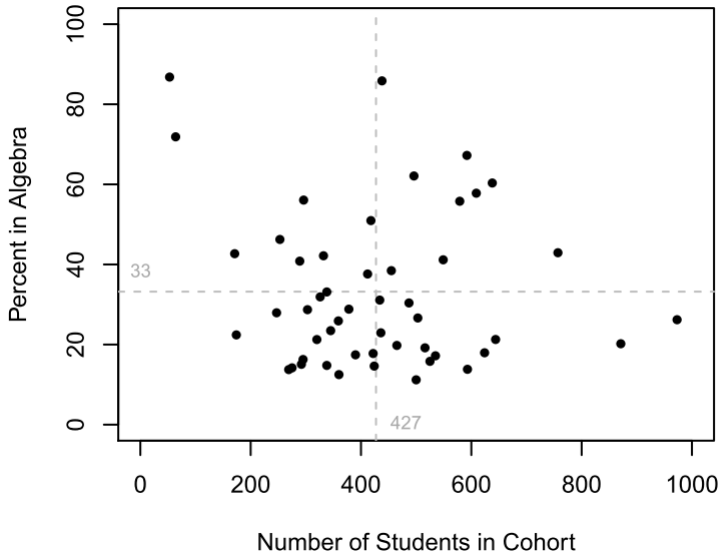

Figure 3.3. Schools for empirical illustration by number of students in cohort and percent of students in Algebra. Grand-means represented by grey dashed lines.

Student-level longitudinal data systematically collected by the school district are available for the cohort from 6th grade through 12th grade (i.e., the 2001-02 school year to the 2007-08 school year). These data include information on student background characteristics, prior academic achievement, course taking records, and high school academic achievement. For

the empirical illustration, background characteristics in 7th grade and academic achievement data in 6th and 7th grade were used as pretreatment covariates (X). I used student performance on the California High School Exit Exam (CAHSEE) mathematics test as the outcome of interest. The CAHSEE mathematics test covers 6th grade, 7th grade, and Algebra 1 mathematics content. Most students take the CAHSEE for the first time in the Spring of their 10th grade year, with additional opportunities to pass the test in later years if necessary. I only examined scale scores from each student's first CAHSEE math attempt (i.e., 10th grade score). By the end of 10th grade, all students in the sample should have had at least one full year of Algebra 1 instruction, and thus exposure to all the mathematics content covered in the test. Their test performance, however, could differ based on the type of mathematics course they were assigned in 8th grade. To simplify the illustration, and focus on the key components of the proposed method, I only examined students with an observed CAHSEE mathematics scale score and key pretreatment covariates.

3.4.2. Methods for empirical illustration

To address research question 4, I compared effect estimation findings from the proposed two-stage matching method to different specifications within the design phase (different propensity score models) and the adjustment phase (no adjustment or one adjustment). I also compared the estimates from the two-stage matching method to alternative, more traditional, effect estimation methods. Specifically, I used an unconditional and conditional multilevel regression model to estimate the grand-mean ATT and between-site ATT variance based on unmatched data, matched data based on a pooled matching method (under different propensity score models), and matched data based on the proposed two-stage matching method.

In addition to addressing research question 4, which includes estimating the extent of school-level effect heterogeneity, I also demonstrated how the proposed method can facilitate exploratory investigations of other effect heterogeneity. This included exploring whether student- and school-level characteristics explain some of the variation in treatment effects. More specific details about the empirical illustration methods are discussed in the following chapter.

3.5. Summary of proposed method and study design

The proposed two-stage matching method consists of three primary phases: (1) a design phase, in which one uses a two-stage matching strategy to construct treatment and control groups that are well balanced along both unit- and site-level key pretreatment covariates; (2) an adjustment phase, in which the observed outcomes for non-local control group matches are adjusted to account for differences in the local and non-local matched control units; and (3) an analysis phase, in which one estimates average causal effects for the treated units and investigates heterogeneity in causal effects through multilevel modeling. The steps in each phase are adapted from Stuart and Rubin (2008) to address a multisite research setting where treatment effect heterogeneity can be examined. In addition to extending the work by Stuart and Rubin, this study compliments the small set of studies that have examined how propensity score model specifications influence effect estimation in multisite settings (Arpino & Mealli, 2011; Su & Cortina, 2009; Thoemmes, 2009; Thoemmes & West, 2011) by testing different propensity score models along with different matching methods.

I used a series of Monte Carlo simulation studies and an empirical illustration to test the proposed method and illustrate its implementation. Chapter 4 is dedicated to the empirical illustration. The main purpose of the empirical illustration was to demonstrate implementation of

the proposed method and further understand how effect estimation based on the proposed method compares to more traditional effect estimation methods (research question 4). Chapter 5 is dedicated to the simulation studies. The main purpose of the simulation studies was to understand how the proposed method performs under different assignment mechanisms, as well as investigate the method's sensitivity to different specifications within each of the three phases. The different simulation conditions are summarized in Table 3.4. In both the empirical illustration and the simulation studies, I conducted all data analyses with the R statistical program (R Development Core Team, 2011). For propensity score matching I relied on the *MatchIt* R package (Ho, Imai, King, & Stuart, 2011) and for multilevel modeling I relied on the *lme4* R package (Bates, Maechler, & Bolker, 2011).

Table 3.4. Summary of conditions tested in the simulation studies.

Phase	Condition Type	Specifications
All three	Treatment assignment mechanism	• Random assignment • Selection on unit-level observables • Selection on unit- and site-level observables • Selection on unit-level observables and site-level observables and unobservables • Selection on unit- and site-level observables and unobservables
Design	Propensity score model	• Single-level logistic regression model • Two-level random intercept (RI) logistic regression model • Two-level random intercept and slope (RIS) logistic regression model
Design	Matching method	• Allow matches within and between sites (pooled matching) • Restrict matching to within-site • Two-stage matching method
Imputation	Number of imputations	• $m = \{0,1,5,10\}$
Analysis	Data preprocessing	• Preprocess data with two-stage matching method • Do not preprocess data
Analysis	Effect estimation model	• Single-level model without covariate adjustment • Single-level model adjusting for observed unit and site covariates • RIS two-level model without covariate adjustment • RIS two-level model adjusting for observed unit-level covariates • RIS two-level model adjusting for observed unit- and site-level covariates

Chapter 4

Empirical Illustration: The Case of Eighth Grade Algebra

Eighth graders in the U.S. are not exposed to as much algebra as their international peers (Schmidt, 2004), and isolated state and school district policy efforts over the last decade have sought to increase the number of students who take their first formal algebra course before high school (Allensworth, Nomi, Montgomery, & Lee, 2009; Burris, Heubert, & Levin, 2006; Clotfelter, Ladd, & Vigdor, 2012; Williams, Haertel, & Kirst, 2011). Past research on selective algebra course placement suggests positive benefits from taking algebra in 8th grade (Gamoran & Hannigan, 2000; Ma, 2005; Smith, 1996), but this research may be limited by overlooked methodological complications. Namely, the strong selectivity of the assignment process may make the regression-based estimates utilized in past research sensitive to parametric assumptions, and between-school heterogeneity in both the assignment process and treatment effects was not explicitly addressed in the past research. In this chapter, I employ the proposed two-stage method to address three substantive questions about the effect of taking algebra in 8th grade:

1. Does assignment to 8th grade algebra affect average student performance on the California High School Exit Exam (CAHSEE)?
2. Does the effect of assignment to 8th grade algebra differ across schools?
3. Are certain factors associated with heterogeneity in the effect of assignment to 8th grade algebra?

As described in the previous chapter, I utilized longitudinal data from a large urban school district to investigate the effect of assignment to 8th grade algebra. To illustrate application of the proposed method, this chapter walks through the three phases of the proposed method—design, adjustment, and analysis—as they pertain to this empirical investigation of 8th grade algebra. The chapter concludes by comparing the average effect estimates based on the proposed method to other standard effect estimation approaches.

4.1. The design phase

As discussed in the previous chapter, the objective in the design phase is to construct a control group that is as similar as possible to the treatment group in terms of the important confounding pre-treatment factors. For the empirical illustration, the important confounding pre-treatment factors fall into three categories: (1) academic performance, (2) academic engagement & program participation, and (3) demographic characteristics. Overall treatment and control group standardized mean differences for the important confounders are presented in Table 4.1. As one might expect, the largest differences were among the academic performance covariates, although some important group differences also existed among the other covariates (e.g., being in a gifted/talented program and being an English learner). An overall indicator of average school achievement is also included in the table. I constructed a school achievement index measure by taking the mean standardized California Standardized Test (CST) scale score in each school across the students' 6th and 7th grade English language arts (ELA) and mathematics CST, after converting each scale score to a z-score. I used the school achievement index to identify site clusters (described below).

Table 4.1. Summary of algebra and pre-algebra differences among important pre-treatment covariates for the original sample and trimmed sample.

	Original Sample				Trimmed Sample			
	PreAlg Mean	Alg Mean	Std Bias	Var Ratio	PreAlg Mean	Alg Mean	Std Bias	Var Ratio
Number of Students	12,733	6,330			12,640	5,518		
Academic Performance								
7th Grade Math CST	291.88	338.44	1.06	1.99	290.98	325.38	0.93	1.13
6th Grade Math CST	288.37	338.15	1.04	2.14	287.58	325.98	0.91	1.25
7th Grade ELA CST	302.00	341.23	0.92	1.36	301.39	333.36	0.79	1.09
6th Grade ELA CST	293.98	330.13	0.90	1.50	293.42	322.63	0.77	1.14
Math GPA	1.64	2.58	0.85	0.96	1.64	2.47	0.76	0.98
Academic Engagement & Program Participation								
Attendance Rate	0.95	0.96	0.28	0.69	0.95	0.96	0.24	0.85
Ever Suspended	0.12	0.07	-0.16	0.64	0.12	0.08	-0.14	0.83
In GATE Program	0.04	0.30	0.80	5.44	0.04	0.23	0.64	2.26
In Magnet Program	0.06	0.14	0.25	1.99	0.06	0.11	0.17	1.29
Student w/Disabilities	0.06	0.01	-0.28	0.23	0.06	0.02	-0.27	0.49
Demographic Characteristics								
English Only	0.24	0.27	0.06	1.07	0.24	0.23	-0.02	0.99
Initial-Fluent EP	0.07	0.10	0.13	1.47	0.07	0.09	0.10	1.16
Reclass.-Fluent EP	0.36	0.49	0.28	1.09	0.36	0.52	0.33	1.04
English Learner	0.33	0.13	-0.49	0.52	0.34	0.15	-0.44	0.76
Free/Reduced Meals	0.73	0.73	0.00	1.00	0.73	0.77	0.07	0.96
New to School	0.04	0.02	-0.11	0.56	0.04	0.02	-0.10	0.76
Female	0.50	0.55	0.10	0.99	0.50	0.56	0.12	0.99
Afr. Am./Black	0.09	0.07	-0.07	0.80	0.09	0.08	-0.06	0.92
Hispanic/Latino	0.78	0.69	-0.19	1.22	0.78	0.75	-0.07	1.05
White	0.09	0.13	0.15	1.45	0.09	0.10	0.05	1.07
Other Race/Ethnicity	0.04	0.10	0.23	2.25	0.04	0.07	0.14	1.32
School Characteristic								
School Ach. Index	0.01	-0.02	-0.07	1.14	0.01	-0.08	-0.25	0.95

Notes: CST = California Standards Test; ELA = English language arts; GPA = grade point average; GATE = gifted and talented education; EP = English proficient.

In this section, I describe the actions I took to preprocess (Ho et al., 2007) the data in the design phase to minimize differences in the treatment and control groups. These actions included trimming the data to remove high achieving treatment students with few similar control students,

identifying site clusters for acceptable between-school matches, estimating the propensity score, executing the two-stage matching method, and assessing covariate balance for the matched sample. Each action is described below.

4.1.1. Trim the data

In most schools, academic performance factors like 7th grade mathematics CST scores differentiate algebra and pre-algebra students. In fact, the 7th grade mathematics CST distributional overlap among algebra and pre-algebra groups is very limited for student who scored in the advanced performance level. Of the 905 students with a 7th grade mathematics CST scale score above 400, only 93 (10%) took pre-algebra in 8th grade. The lack of available control students at the advanced level of 7th grade math performance raises concerns for both regression-based and matching-based adjustment methods. From a regression perspective, the lack of covariate overlap means treatment effect estimates pertaining to advanced students requires extrapolation to a region with little data support. From a matching perspective, the lack of overlap means it will be very difficult to find "similar" control student matches for treatment students in this advanced region.

To avoid these concerns, I restricted the analysis to students who scored below 400 on the 7th grade mathematics CST (i.e., students who did not score in the "advanced" performance level). Trimming the data in this way excluded 812 (13%) algebra students, but facilitated more plausible estimation for the remaining 5,330 algebra students. Trimming the data did not dramatically alter the overall treatment group student population, and, while reduced in most cases, the important covariate group differences still remained (see Table 4.1). One change to note, however, is that the school achievement index difference increased between algebra and

pre-algebra groups after trimming the sample. This suggests that the algebra students excluded from the analysis tended to reside in higher achieving schools.

4.1.2. Identify site clusters

For the two-stage matching, between-site matches are only accepted for sites within the same cluster of "similar" sites. In practice, one could take a variety of different approaches to define site clusters. For example, clusters could be constructed based on a multivariate array of site characteristics, based on a single key covariate, and/or based on geographic location. For simplicity of demonstration, I defined site clusters based on each school's average achievement index defined above. Using this index, I grouped schools into one of five clusters based on the index quintile ranges (see Figure 4.1). While many schools in quintiles 2-4 had similar average achievement to schools in neighboring quintiles, the main benefit of defining site clusters is likely the prevention of matches between, for example, schools in quintile 1 and quintile 5, where the large differences in math achievement are likely to reflect significant differences in school climate/context.

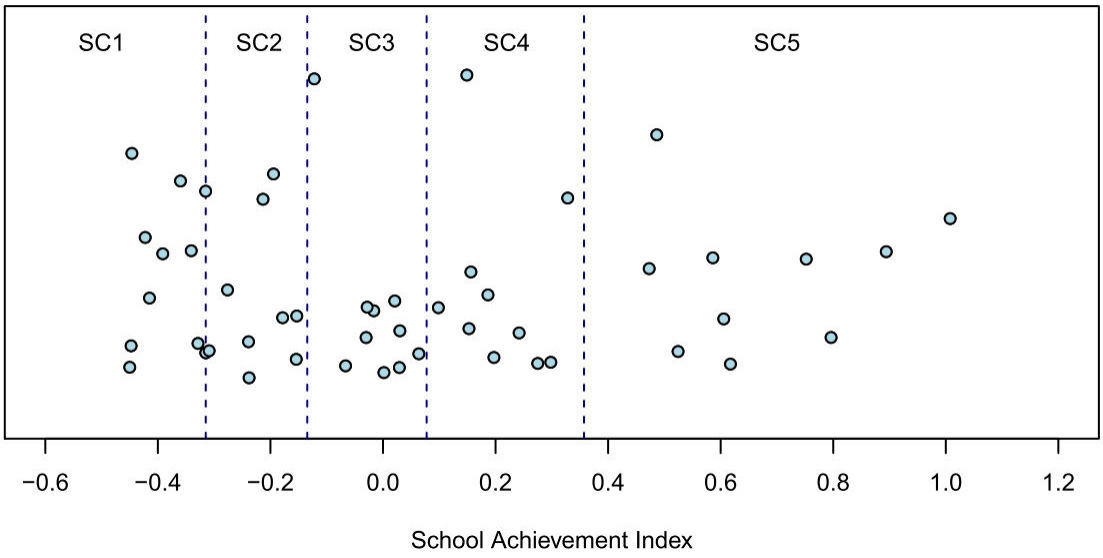

Figure 4.1. Distribution of schools within the five site clusters (SC) defined by the school achievement index.

4.1.3. Estimate each unit's propensity for 8th grade algebra

Given the desire to match on multiple covariates and allow the relative importance of unit-level covariates to vary across sites, a random-intercept-and-slope (RIS) multilevel logistic regression model was employed to estimate each unit's propensity for assignment to 8th grade algebra. The model has the following general structure:

$$\text{Level 1: } \ln\left(\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right) = \beta_{0j} + \boldsymbol{\beta}_{1j}\mathbf{X}_{ij}^{gp} + \boldsymbol{\beta}_{2j}\mathbf{Z}_{ij}^{gp}, \tag{4.1}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \boldsymbol{\gamma}_{01}\mathbf{S}_{j}^{gd} + u_{0j},$$

$$\boldsymbol{\beta}_{1j} = \boldsymbol{\gamma}_{10} + \mathbf{u}_{1j},$$

$$\boldsymbol{\beta}_{2j} = \boldsymbol{\gamma}_{10},$$

where level-1 includes a vector of student-level covariates (**X**) with coefficients allowed to vary across schools and a vector of student-level covariates (**Z**) with coefficients fixed across schools.

All the level-1 covariates are centered on their respective group mean. At level-2, a series of site-cluster indicators (**S**) are included to capture possible between-cluster differences.

I made specific model-specification decisions regarding covariate inclusion/exclusion, covariate transformations (including interactions and quadratic/cubic terms), and which covariates should have random coefficients based on exploratory model estimation and prior research on the topic of 8th grade algebra assignment (Rickles, 2011). All the covariates listed in Table 4.1 were included in the final propensity score model in some form. Model-specification features include:

- An interaction between 7th grade math CST scale score and math GPA;
- Quadratic and cubic terms for 7th grade math CST and GPA;
- Random slopes for 7th grade math CST and GPA, and their interaction term;
- Dichotomous performance level indicators for 6th grade math CST instead of the semi-continuous scale score measure;
- Dichotomous course grade indicators for the second semester of 7th grade math in addition to the math GPA measure.

The fixed effect parameter estimates from the final estimated model are presented in Table 4.2, while the random effect estimates are presented in Table 4.3.

Table 4.2. Propensity score model fixed effect estimates.

Fixed Effect	Estimate	Std. Err.	Z-Value	P-Value
Intercept	-1.894	0.501	-3.782	0.000
Academic Performance				
7th Grade Math CST	0.853	0.106	8.076	0.000
7th Grade CST-Squared	-0.029	0.037	-0.795	0.427
7th Grade CST-Cubed	0.008	0.023	0.374	0.708
6th Grade CST PL: FBB	-0.615	0.113	-5.441	0.000
6th Grade CST PL: BB	-0.455	0.067	-6.820	0.000
6th Grade CST PL: PP	0.744	0.084	8.870	0.000
Math GPA	0.559	0.126	4.454	0.000
Math GPA-Squared	-0.066	0.044	-1.491	0.136
Math GPA-Cubed	-0.054	0.028	-1.908	0.056
7th Grade Math Grade: A	0.599	0.132	4.552	0.000
7th Grade Math Grade: B	0.296	0.081	3.632	0.000
7th Grade Math Grade: D	-0.277	0.085	-3.281	0.001
7th Grade Math Grade: F	-0.527	0.143	-3.697	0.000
Interaction: CST*GPA	0.005	0.061	0.081	0.935
7th Grade ELA CST	0.234	0.047	4.994	0.000
6th Grade ELA CST	0.392	0.047	8.326	0.000
Academic Engagement & Program Participation				
7th Grade Attendance Rate	0.096	0.027	3.522	0.000
Ever Suspended	-0.026	0.085	-0.306	0.760
In GATE Program	1.472	0.093	15.898	0.000
In Magnet Program	-0.030	0.100	-0.294	0.769
Student w/Disabilities	-0.566	0.156	-3.628	0.000
Demographic Characteristics				
English Learner	-0.473	0.105	-4.523	0.000
Initial-Fluent EP	0.137	0.110	1.247	0.212
Reclass.-Fluent EP	-0.083	0.087	-0.957	0.339
Free/Reduced Meals Participant	0.200	0.063	3.158	0.002
New to School in 8th Grade	-0.539	0.145	-3.726	0.000
Female	0.084	0.050	1.703	0.089
Afr. Am./Black	-0.098	0.138	-0.708	0.479
Hispanic/Latino	-0.318	0.109	-2.920	0.004
Other Race/Ethnicity	0.192	0.136	1.418	0.156
Site Cluster Effects				
Site Cluster 1	0.693	0.663	1.044	0.296
Site Cluster 2	0.147	0.662	0.221	0.825
Site Cluster 4	-0.482	0.673	-0.717	0.474
Site Cluster 5	-0.150	0.674	-0.223	0.824

Table 4.3. Propensity score model random effect estimates.

Random Effect	Variance	Std. Dev.	Correlation Matrix			
Intercept	3.478	1.865	1.000			
7th Grade Math CST	0.324	0.569	-0.634	1.000		
Math GPA	0.376	0.613	-0.130	-0.039	1.000	
Interaction: CST*GPA	0.086	0.294	0.360	-0.587	-0.207	1.000

The estimated model parameters indicate that students with relatively high predicted propensity scores were more likely to exhibit higher academic performance in 6th and 7th grade, higher school day attendance, and participate in GATE. Conversely, students with relatively low propensity scores were more likely to be a student with disabilities, an English learner, or attend a new school. The between-school parameter variance estimates indicate that the magnitude of a student's predicted propensity score depended on the school a student attended, with the standard deviations of the random effect parameters roughly of equal magnitude as their respective grand-mean point estimates. Additionally, the importance of 7th grade math CST scale scores in predicting a student's assignment to 8th grade algebra in a given school was negatively related to the overall average propensity of algebra placement in a school ($r_{intercept,CST} = -0.634$). In other words, CST scores were less of a placement criterion in schools with higher percentages of students in algebra. Note, however, that a much weaker relationship existed between overall average propensity score and emphasis placed on math GPA ($r_{intercept,GPA} = -0.130$). These relationships are depicted in Figure 4.2.

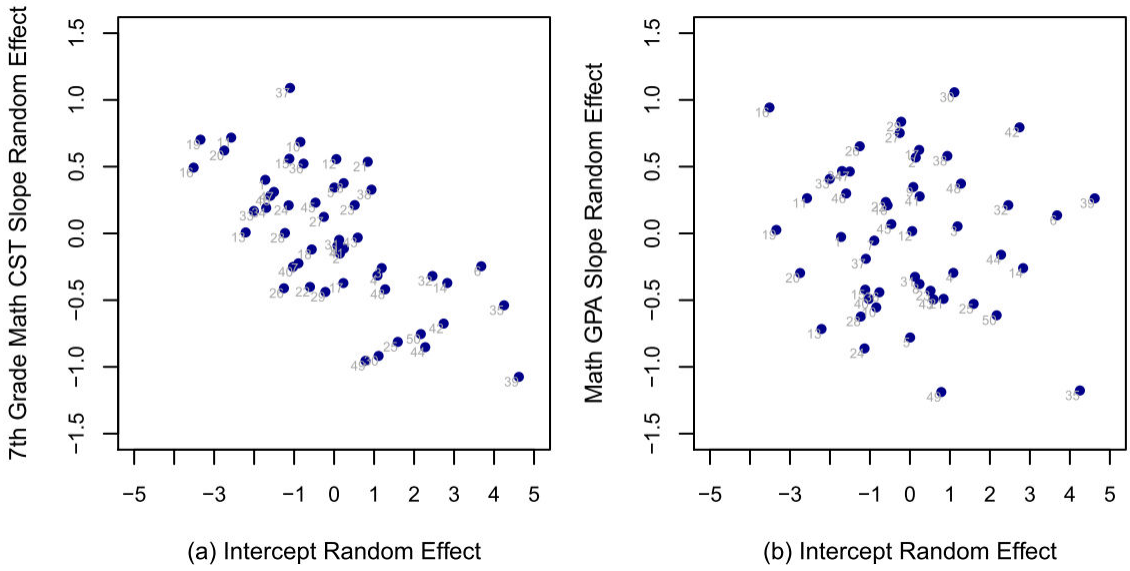

Figure 4.2. Relationship in propensity score model school-level intercept random effect and slope random effect estimates.

Each student's predicted, or estimated, probability of 8th grade algebra assignment was generated from the estimated propensity score model. The estimated log-odds propensity score distributions for both the algebra and pre-algebra groups are displayed in Figure 4.3. The limited overlap in the propensity score distributions reiterates the fact that algebra and pre-algebra students are very different entering 8th grade. As a result, one's ability to find "quality" matches for algebra students with high propensity scores may be limited. Furthermore, the degree of propensity score overlap within schools differed across schools (see Figure 4.4), with very limited overlap in some schools. (e.g., schools 6 and 35).

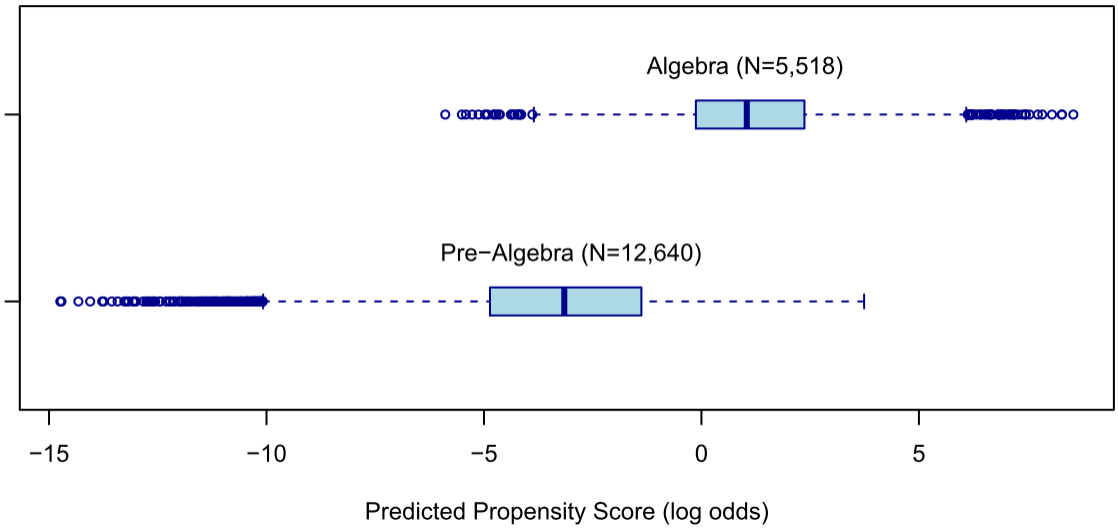

Figure 4.3. Predicted propensity score distributions for algebra and pre-algebra groups.

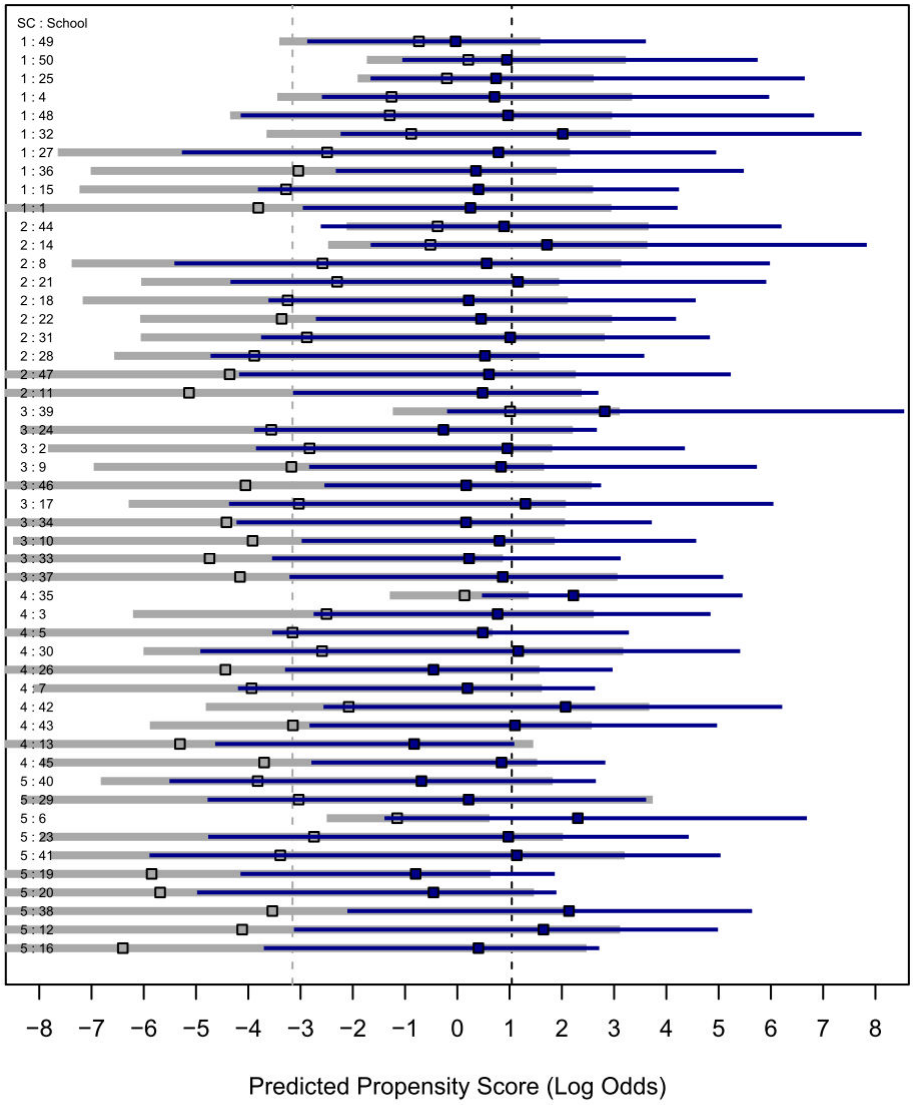

Figure 4.4. Within-school predicted propensity score distributions for algebra (dark) and pre-algebra (grey) groups before matching. Boxes represent median value and horizontal lines represent the min-max range. Vertical dashed lines mark the overall mean for each group.

4.1.4. Execute two-stage matching and assess resulting matched sample

The two-stage matching strategy was implemented separately for each school following the steps outlined in the previous chapter. The matching criteria were one-to-one within caliper (0.25 standard deviation) matching based on the log-odds propensity score combined with the

Mahalanobis distance based on 7th grade mathematics CST scale score and 7th grade math GPA. The resulting matched data set included 5,269 (95%) of the 5,518 algebra students, which means acceptable matches were not found for 249 algebra students. Overall, only 57% of the matched treatment students were matched to a control student within the same school. So moving from a within-school matching approach to the two-stage matching approach retained a substantial number of treatment students in the analysis.

It is important to note, however, that in order to retain more treatment students with the between-school matching, some control students had to be matched to multiple treatment students. For each school, the matching (both within-school and between-school matching) was conducted without replacement, i.e., control students were only matched to one treatment student. Pooling the matched data across schools, however, can result in duplication of control student records. In the pooled matched data for this analysis, 3,597 (68%) of the 5,269 control student records represented unique pre-algebra students. Most of these control students (2,730) were only matched to one treatment student, while 501 students were matched to two treatment students and 366 were matched to more than two treatment students. At the extreme end, four control students were matched to ten different treatment units. Since control students are not duplicated within a given matched school, repeated observations should not affect within-school estimates. Assessing the extent to which control unit duplication influences between-school estimates, and methods for adjusting for any potential bias, is a topic for future research.

The degree to which a within-school match, or any match, was found for the treatment students differed across schools. Figure 4.5 displays the proportion of treatment students matched for each school and the proportion of matches found within the same school. In some schools, less than 90% of the treatment units had an acceptable match (e.g., school 31).

Furthermore, in some schools (e.g., school 35) very few treatment students were matched to control students within the same school, but acceptable matches were available for most of the students when going to other schools for matches. Without the two-stage matching, schools with very few within-school matches would contribute little to average effect estimation.

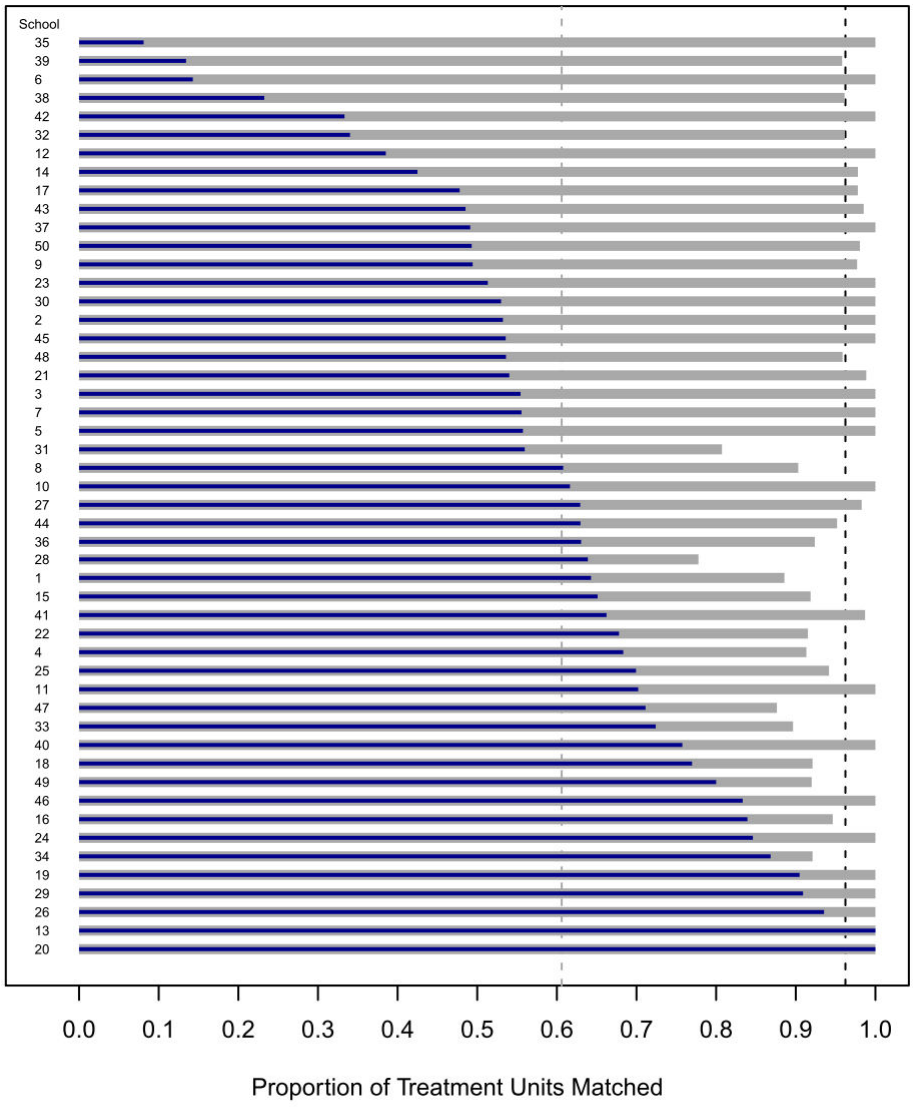

Figure 4.5. Total proportion of treatment students matched (grey bar) and proportion of within-school matches (dark bar), by school. Vertical dashed lines mark the overall mean for each group.

Given the degree of between-school heterogeneity in the propensity score parameters documented in Table 4.3 and Figure 4.2, conducting between-school matching based on the predicted propensity score derived from the target treatment school's parameter estimates seems particularly salient. Changing the propensity score used for each between-school match makes matching diagnostics based on the original predicted propensity score less meaningful, however. Nevertheless, comparing the propensity score distribution before and after matching provides a rough summary about matching performance across the multivariate distribution of observed covariates. Figure 4.6 displays the degree of propensity score overlap within each school after matching. For most schools, the distributional overlap in the propensity score was dramatically improved relative to the pre-matched sampled (see Figure 4.4). For some schools (e.g., schools 35 and 39), however, the distributional overlap remained extremely limited. The schools with limited predicted propensity score overlap were also the schools with relatively few within-school matches.

Covariate-specific balance diagnostics provide a better test of matching performance than generalizations based on the propensity score. Overall balance statistics for each of the key covariates are presented in Table 4.4 based on the unmatched sample and the matched sample. Matching dramatically reduced the overall average covariate differences between the treatment and control groups. Prior to matching, for example, algebra students scored, on average, over half a standard deviation higher on academic performance measures than the pre-algebra students. After restricting the sample to matched algebra and pre-algebra students, the average differences were less than 0.20 of a standard deviation. Across all the important pre-treatment covariates, the standardized mean difference for 11 covariates was greater than 0.20 for the

unmatched sample, but for the matched sample no standardized mean difference was greater than 0.20 of a standard deviation.

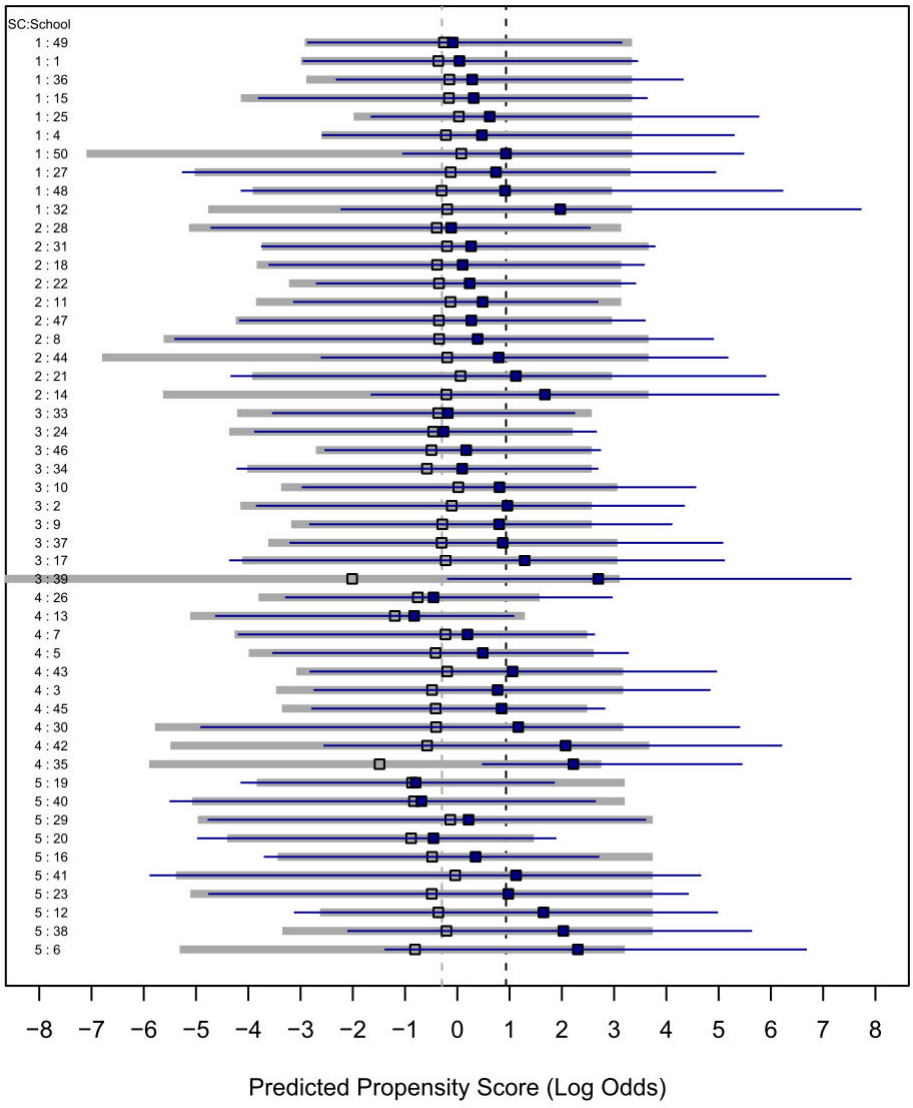

Figure 4.6. Within-school predicted propensity score distributions for algebra (dark) and pre-algebra (grey) groups after matching. Boxes represent median value and horizontal lines represent the min-max range. Vertical dashed lines mark the overall mean for each group.

Table 4.4. Summary of algebra and pre-algebra differences among important pre-treatment covariates for the unmatched and matched samples.

	Unmatched Sample				Matched Sample			
	PreAlg Mean	Alg Mean	Std Bias	Var Ratio	PreAlg Mean	Alg Mean	Std Bias	Var Ratio
Number of Students	12,640	5,518			5,269	5,269		
Academic Performance								
7th Grade Math CST	290.98	325.38	0.93	1.13	320.72	323.40	0.07	1.01
6th Grade Math CST	287.58	325.98	0.91	1.25	317.78	323.50	0.14	1.03
7th Grade ELA CST	301.39	333.36	0.79	1.09	327.55	331.34	0.09	0.98
6th Grade ELA CST	293.42	322.63	0.77	1.14	317.03	320.63	0.09	1.02
Math GPA	1.64	2.47	0.76	0.98	2.38	2.44	0.05	1.00
Academic Engagement & Program Participation								
Attendance Rate	0.95	0.96	0.24	0.85	0.96	0.96	0.03	1.00
Ever Suspended	0.12	0.08	-0.14	0.83	0.08	0.08	-0.02	0.97
In GATE Program	0.04	0.23	0.64	2.26	0.16	0.20	0.15	1.11
In Magnet Program	0.06	0.11	0.17	1.29	0.10	0.11	0.02	1.02
Student w/Disabilities	0.06	0.02	-0.27	0.49	0.02	0.02	-0.03	0.86
Demographic Characteristics								
English Only	0.24	0.23	-0.02	0.99	0.23	0.23	0.01	1.00
Initial-Fluent EP	0.07	0.09	0.10	1.16	0.08	0.09	0.04	1.05
Reclass.-Fluent EP	0.36	0.52	0.33	1.04	0.50	0.51	0.03	1.00
English Learner	0.34	0.15	-0.44	0.76	0.19	0.16	-0.06	0.94
Free/Reduced Meals	0.73	0.77	0.07	0.96	0.77	0.76	-0.01	1.01
Non-Resident School	0.21	0.20	-0.03	0.98	0.20	0.20	0.01	1.01
New to School	0.04	0.02	-0.10	0.76	0.03	0.03	-0.02	0.94
Over-Age for 8th Grade	0.04	0.02	-0.09	0.79	0.02	0.03	0.02	1.08
Female	0.50	0.56	0.12	0.99	0.54	0.56	0.04	1.00
Afr. Am./Black	0.09	0.08	-0.06	0.92	0.08	0.08	0.00	1.00
Hispanic/Latino	0.78	0.75	-0.07	1.05	0.75	0.75	0.00	1.00
White	0.09	0.10	0.05	1.07	0.10	0.10	-0.01	0.99
Other Race/Ethnicity	0.04	0.07	0.14	1.32	0.07	0.07	0.02	1.02
School Characteristic								
School Ach. Index	0.01	-0.08	-0.25	0.95	-0.07	-0.08	-0.02	1.00

Even after matching, two covariates were still of some concern because their standardized mean difference was over 0.10: 6th grade math CST scale score and GATE. It is interesting that, overall, the standardized mean covariate differences were smaller for the

between-school matched students than the within-school matched students (see Table 4.5). This was particularly true for the academic performance indicators, where the mean differences for the within-school matches were a little over 0.10 of a standard deviation, but around 0.05 of a standard deviation for the between-school matches. This likely reflects the trade-off that comes with within-site matching, where both site-level observed and unobserved factors are equated between groups at the possible expense of minimizing differences in unit-level factors.

The reduction in covariate standardized mean differences from matching generally held within each school. For example, Figure 4.7 displays the 7th grade math CST scale score distribution for algebra (dark) and pre-algebra (grey) students within each school before matching (left panel) and after matching (right panel). For this key covariate, the distributional overlap was dramatically improved within each of the 50 schools after matching. This was even true in schools like 35 and 39, where the propensity score overlap was still limited after matching. Even after matching, however, some schools maintained sizable group differences for certain potentially important covariates. The min-max range and median within-school absolute standardized mean difference for each of the important covariates is displayed in Figure 4.8, based on both the unmatched and matched samples. For all the covariates, the group differences in at least half of the schools were less than 0.2 of a standard deviation after matching, but larger differences for all the covariates remained in at least one school.

Table 4.5. Summary of algebra and pre-algebra differences among important pre-treatment covariates for the within-school and between-school matched samples.

	Within-School Matched Sample				Between-School Matched Sample			
	PreAlg Mean	Alg Mean	Std Bias	Var Ratio	PreAlg Mean	Alg Mean	Std Bias	Var Ratio
Number of Students	3,012	3,012			2,257	2,257		
Academic Performance								
7th Grade Math CST	309.78	313.92	0.11	1.03	335.32	336.05	0.02	1.02
6th Grade Math CST	304.44	312.57	0.19	1.06	335.59	338.08	0.06	1.05
7th Grade ELA CST	316.54	321.94	0.13	1.00	342.25	343.87	0.04	0.97
6th Grade ELA CST	307.28	312.09	0.13	1.02	330.05	332.03	0.05	1.05
Math GPA	2.14	2.23	0.09	1.01	2.72	2.72	0.00	1.02
Academic Engagement & Program Participation								
Attendance Rate	0.96	0.96	0.04	1.00	0.96	0.96	0.03	1.00
Ever Suspended	0.09	0.09	0.00	1.01	0.08	0.06	-0.05	0.92
In GATE Program	0.09	0.12	0.11	1.16	0.25	0.31	0.20	1.07
In Magnet Program	0.09	0.09	0.02	1.03	0.12	0.12	0.01	1.01
Student w/Disabilities	0.03	0.02	-0.03	0.90	0.02	0.01	-0.03	0.77
Demographic Characteristics								
English Only	0.22	0.22	0.00	1.00	0.25	0.26	0.02	1.01
Initial-Fluent EP	0.07	0.08	0.04	1.07	0.10	0.11	0.03	1.04
Reclass.-Fluent EP	0.46	0.49	0.06	1.00	0.55	0.54	-0.02	1.00
English Learner	0.25	0.21	-0.09	0.94	0.11	0.10	-0.02	0.96
Free/Reduced Meals	0.77	0.77	0.01	1.00	0.77	0.76	-0.03	1.02
Non-Resident School	0.18	0.19	0.02	1.01	0.22	0.22	-0.01	1.00
New to School	0.03	0.03	0.00	0.99	0.03	0.02	-0.04	0.85
Over-Age for 8th Grade	0.03	0.03	0.01	1.04	0.01	0.02	0.03	1.18
Female	0.53	0.54	0.02	1.00	0.54	0.57	0.07	0.99
Afr. Am./Black	0.08	0.08	-0.02	0.98	0.07	0.08	0.03	1.05
Hispanic/Latino	0.78	0.77	-0.02	1.01	0.72	0.72	0.02	0.99
White	0.08	0.09	0.02	1.03	0.13	0.12	-0.05	0.95
Other Race/Ethnicity	0.06	0.06	0.03	1.05	0.09	0.09	0.00	1.00
School Characteristic								
School Ach. Index	-0.10	-0.10	0.00	1.00	-0.03	-0.05	-0.05	1.01

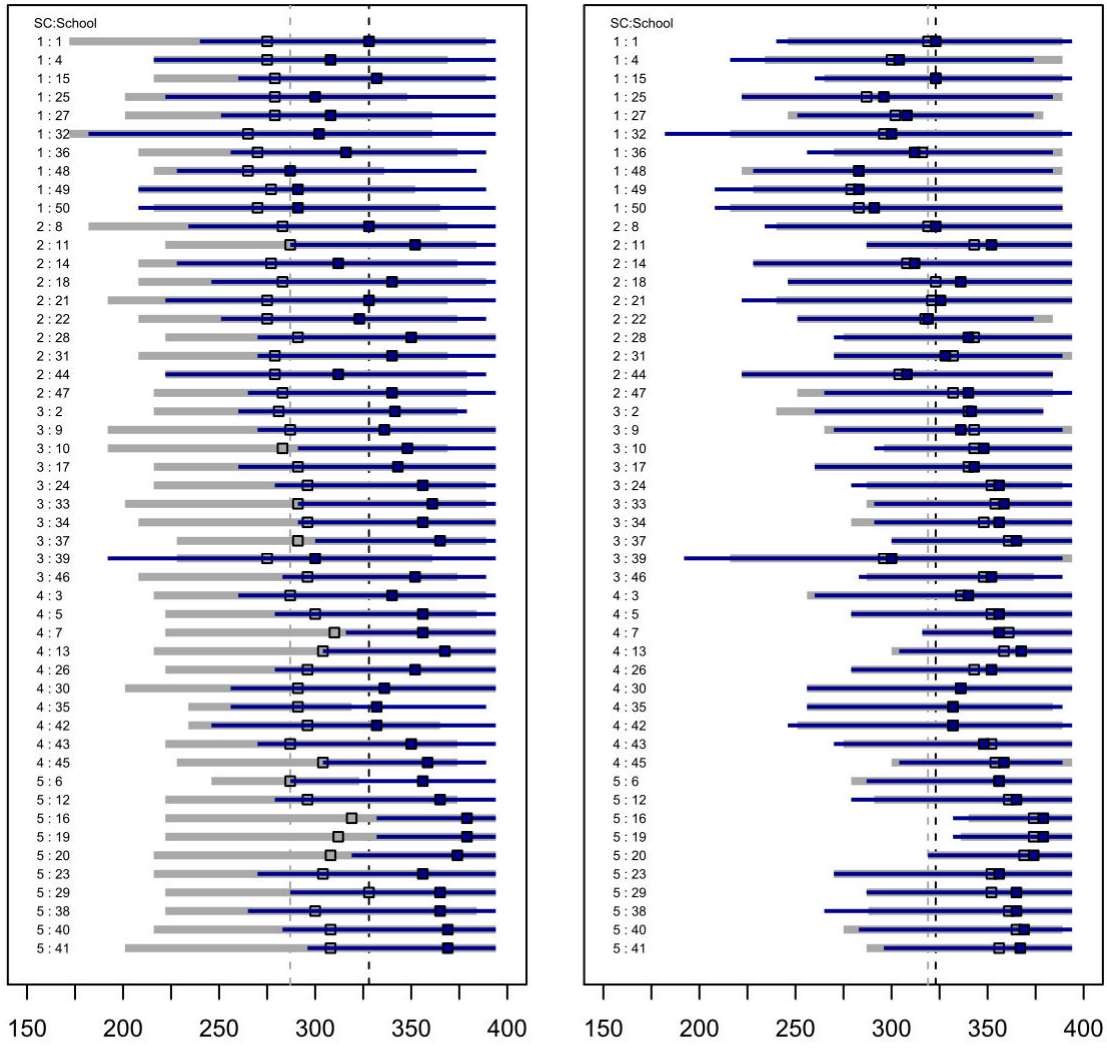

Figure 4.7. Within-school 7th grade mathematics CST scale score distributions for algebra (dark) and pre-algebra (grey) groups before (a) and after (b) matching. Boxes represent median value and horizontal lines represent the min-max range. Vertical dashed lines mark the overall mean for each group.

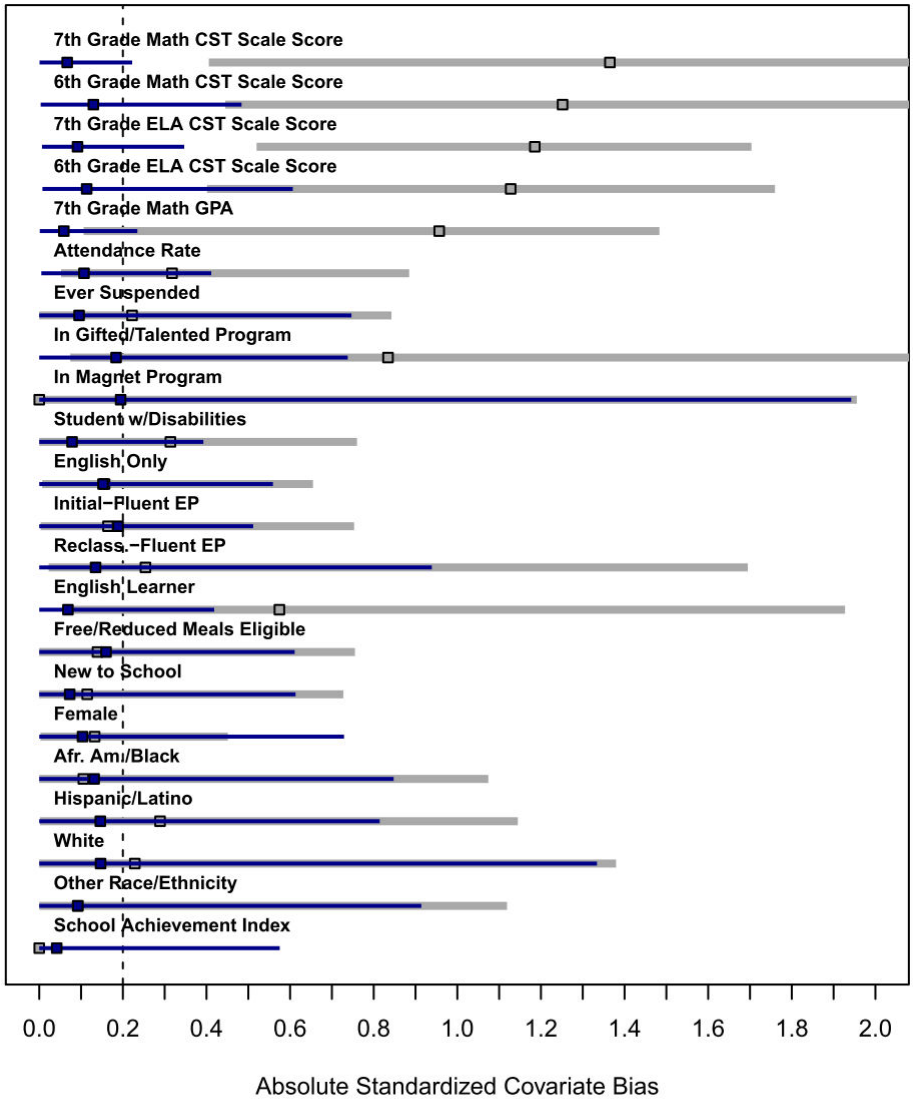

Figure 4.8. Within-school absolute standardized mean differences for important covariates before matching (grey) and after matching (dark). Boxes represent median value and horizontal lines represent the min-max range.

One concerning within-school difference that remained after matching has to do with whether a student was in a magnet program or not. In a handful of schools, the standardized mean difference for the matched sample was larger than one standard deviation, and the median within-school difference actually increased after matching. Since the magnet program covariate

received relatively little weight in the propensity score estimation, it is possible that finding acceptable matches for treatment students in some schools based on the propensity score required a trade-off between improved covariate balance on heavily weighted covariates (e.g., academic performance covariates) and worsened balance on lower weighted covariates like magnet program. Regression-based covariate adjustments in the analysis phase may be necessary to take these post-matching within-school differences into account when estimating average treatment effects, even though the overall mean group differences for the covariates were not large.

4.2. The adjustment phase

Since 43% of the matched treatment students were matched to a control student in another school, differences in school-level factors may bias treatment effect estimation. The objective in the adjustment phase is to estimate the extent of possible school-level bias and adjust the outcomes of matched between-school control students to account for this bias. In this section, I describe the actions I took in the adjustment phase to make corrections in the control student outcomes. These actions—following the methods outlined in the previous chapter—included estimating school effects, multiply imputing plausible school effect values, and adjusting the observed outcome based on these values. Each action is described below.

4.2.1. Estimate school effects

I estimated school effects using a subsample of the control unit data, where all control units (before matching) with a predicted propensity score within the same range as the matched control units were included in the estimation. For the full unmatched sample of control students

(N=12,640), the predicted propensity score ranged from -14.74 to 3.74, with a median value of -3.15. When the sample was restricted to students with predicted propensity scores within the range of matched control units (N=12,358), the predicted propensity score ranged from -9.00 to 3.74, with a median value of -3.09. Subsampling in this way strikes a balance between retaining enough units in the data to estimate school effects with some precision and ensuring the estimates are based on units comparable to the units targeted for adjustment.

I generated parameter estimates and empirical Bayes (EB) point estimates for the school effects following Equation 3.2. To allow variation in school effects across levels of student prior achievement, 7th grade mathematics CST scale score was included as a level-1 independent variable. The CST score was centered on the grand-mean of the student's respective site cluster so the estimated contextual school effects are relative to each site cluster. The CST scale score was also normalized by the overall sample standard deviation for easier interpretation of the coefficient estimate. The model estimated parameters are reported in Table 4.6. The intercept estimate represents the average control student outcome for a student in site cluster 3 (the omitted site cluster category), while the site cluster estimates indicate the estimated average difference in control student outcomes in the other site clusters. Similarly, the CST coefficient estimate indicates the expected change in the outcome in site cluster 3 for each standard deviation change in 7th grade math achievement, while the cross-level interactions with the site cluster indicators how that estimated relationship differs across site clusters.

Table 4.6. School effects model parameter estimates.

Fixed Effect	Estimate	Std. Err.	Z-Value
Intercept	359.43	1.51	237.89
Site Cluster 1	-13.48	2.13	-6.34
Site Cluster 2	-7.93	2.12	-3.74
Site Cluster 4	4.43	2.18	2.03
Site Cluster 5	14.98	2.16	6.93
7th Grade Math CST	18.31	0.84	21.85
Interaction: CST*SC1	-2.27	1.20	-1.89
Interaction: CST*SC2	-0.69	1.15	-0.59
Interaction: CST*SC4	1.98	1.21	1.64
Interaction: CST*SC5	0.33	1.19	0.28

Random Effects	Variance	Std.Dev.	Cor.
Intercept	20.53	4.53	
7th Grade Math CST	4.01	2.00	-0.04
Residual	406.56	20.16	

The random effects for both the intercept and CST indicate that even after accounting for differences across site clusters, meaningful between-school variation existed in terms of the average control group outcome and the relationship between prior achievement and the outcome. The point estimate and approximate 95% confidence interval for each school's EB random effects are displayed in Figure 4.9. The left panel shows the degree to which school effects for an average site cluster control student differed across schools. In site cluster 1, for example, control students with average 7th grade CST math achievement performed substantially better on the CAHSEE in school 4 relative to the other schools in the site cluster. The right panel shows the degree to which the CST slope differed across schools within the same site cluster. For example, the slope for school 4 was similar to the average slope for site cluster 1, while the slope for school 15 was steeper relative to the site cluster average. This suggests that while the average

site cluster 1 student was expected to score higher on the CAHSEE if in school 4 versus school 15, the expected difference was larger for relatively low achieving students and smaller for relatively high achieving students. The adjustment for between-school matched control students takes these differential school effects into account.

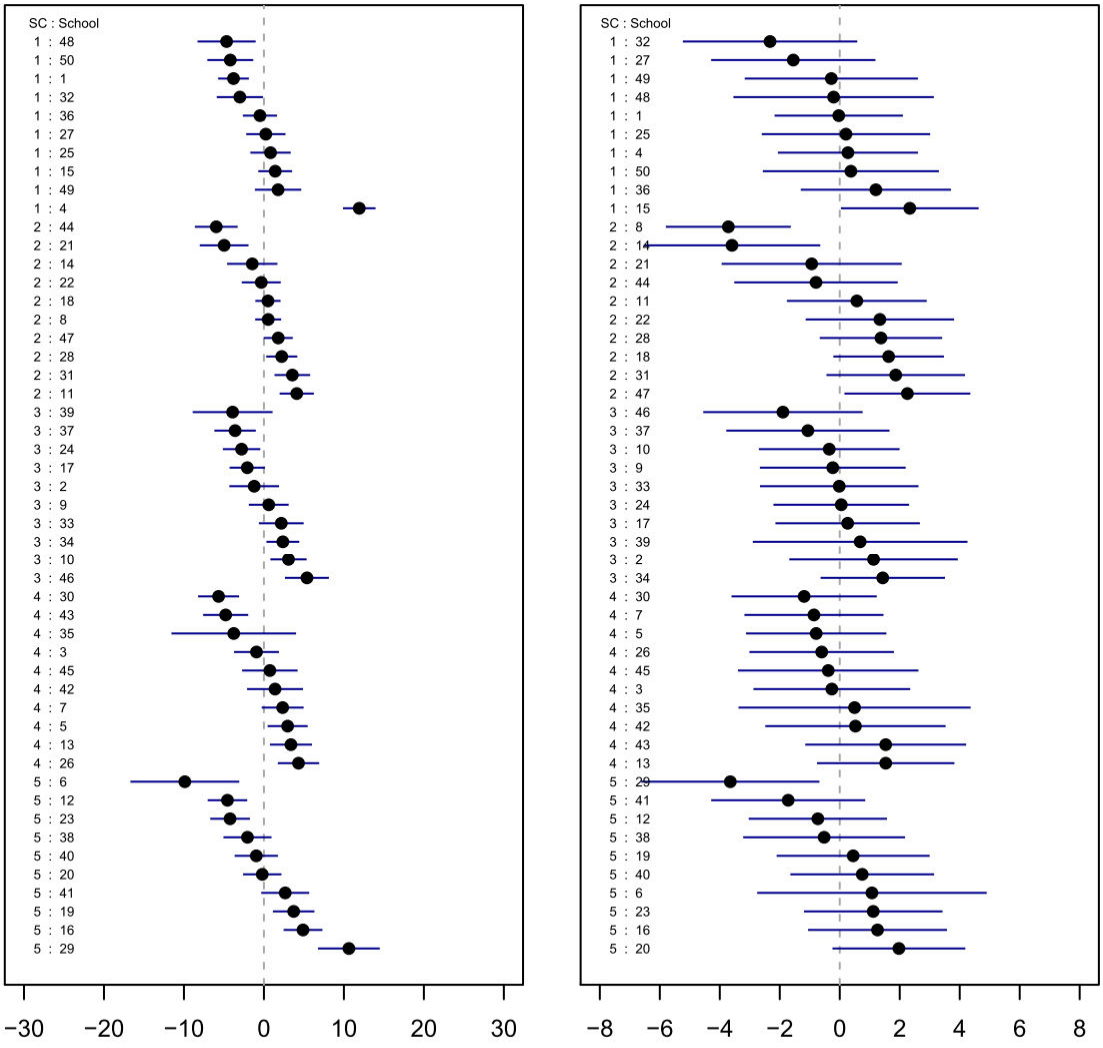

Figure 4.9. School empirical Bayes random effect point estimates and approximate 95% confidence intervals for the school effects model intercept and slope.

4.2.2. Multiply impute plausible school effect values

Figure 4.9 also indicates uncertainty in the school effect point estimates. To incorporate this uncertainty in the analysis, five plausible intercept and slope random effects were imputed for each school following the methods described in the previous chapter. The first step in the imputation process was to draw a sample tau matrix from an inverse-Wishart distribution given the following model estimated tau matrix and 44 degrees of freedom:

$$\hat{\mathbf{T}} = \begin{bmatrix} 20.53 & -0.36 \\ -0.36 & 4.01 \end{bmatrix}$$

This distribution is visually represented in Figure 4.10 based on 1,000 samples to demonstrate how the distribution is centered on the model estimated variance components with right-skewed density around the estimated values. Given a sampled tau matrix and the model estimated residual variance, each school's random effect variance matrix was re-calculated and used to draw a plausible intercept and slope random effect from a multivariate normal distribution centered on the model estimated random effects for each school. Repeating this process for each of the five imputations resulted in five plausible random effect intercept and slope values for each school, as displayed in Figure 4.11.

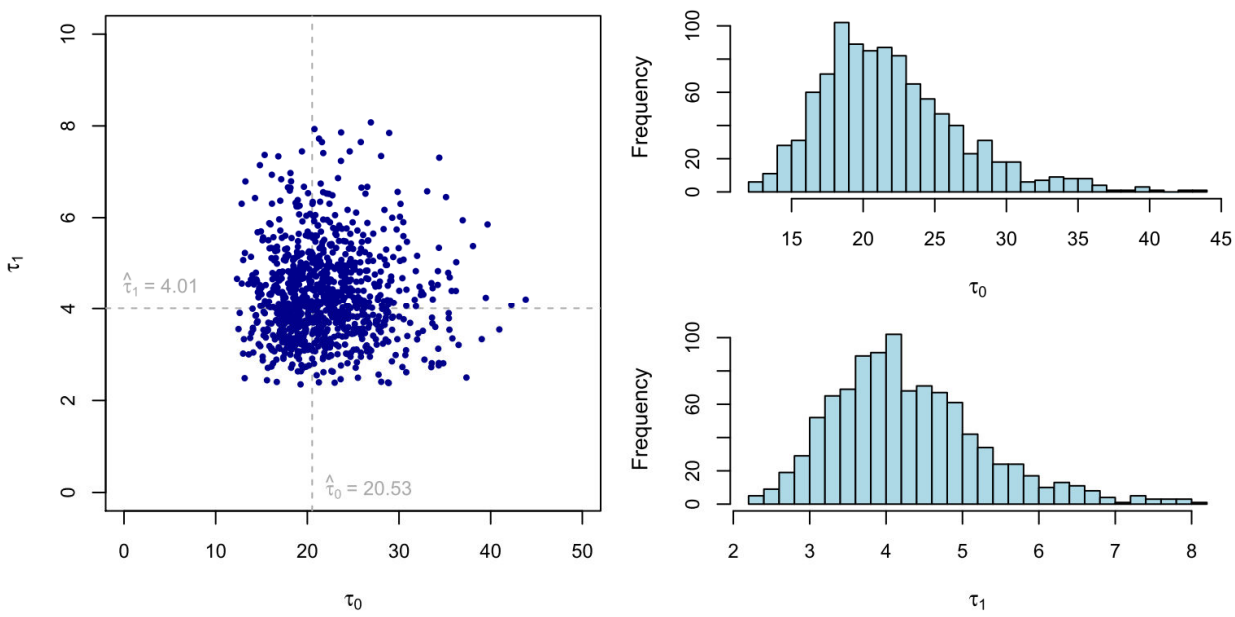

Figure 4.10. Bivariate and univariate distributions of the variance components based on 1,000 random draws from the inverse-Wishart distribution.

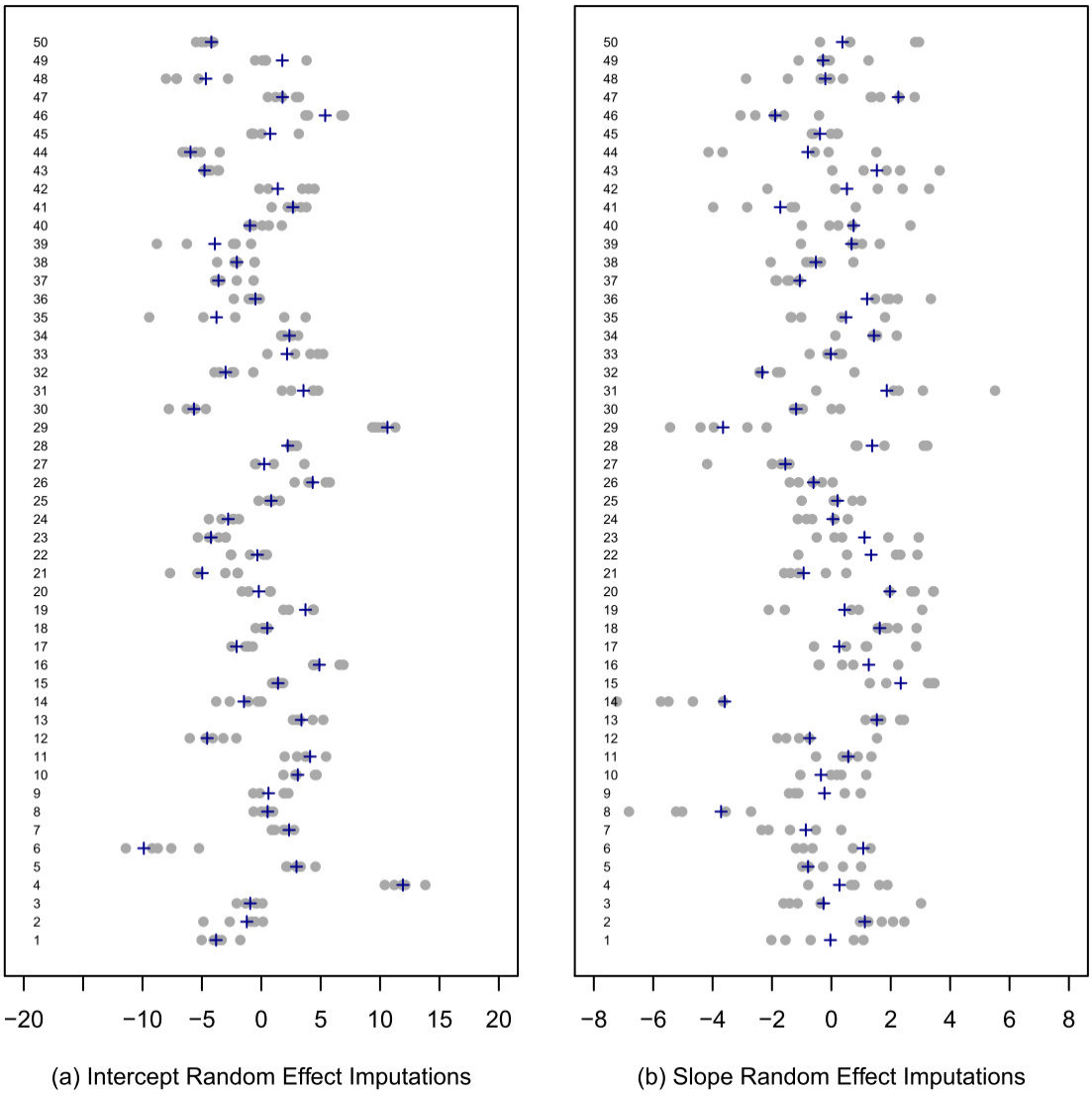

Figure 4.11. Imputed intercept and slope random effects (grey circles) and model estimated random effects (+) for each school.

4.2.3. Adjust observed outcome values

Each of the five imputed plausible school effect values was merged with the matched student data to create five different data sets for analysis. For the between-school matched control students, each data set included the sampled school effect values for the student's school

and for the school of the treatment student s/he was matched to. The difference in these school effects was used to adjust the observed CAHSEE scale score value for the between-school matched control students following Equation 3.3. In four cases, the adjustment resulted in a scale score slightly above the highest possible score, 450, on the CAHSEE. In these cases, the adjusted score was trimmed to 450. Table 4.7 provides an example of the adjustments made for three control students matched to students in school 1. In each case, the control student school was higher achieving, on average, than school 1, so the observed outcome for each control student was adjusted downward. By adjusting the observed outcome in this way, I attempted to extract possible school-level bias induced by the between-school matching from the analysis.

Table 4.7. Example outcome adjustment for three control students matched to treatment students in school 1.

Student ID	School ID	Observed Outcome ($Y(0)$)	CST^c	m	$u^*_{0j=1}$	$u^*_{0j'}$	$u^*_{1j=1}$	$u^*_{1j'}$	Adjustment	Adjusted Outcome ($\tilde{Y}(0)$)
125666	4	401	1.62	1	-3.62	10.40	-1.55	1.61	-19.14	381.86
125666	4	401	1.62	2	-4.01	13.81	-2.02	0.78	-22.36	378.64
125666	4	401	1.62	3	-5.02	12.13	0.76	-0.78	-14.66	386.34
125666	4	401	1.62	4	-1.77	11.21	-0.70	1.89	-17.17	383.83
125666	4	401	1.62	5	-3.34	11.87	1.08	0.66	-14.53	386.47
108589	15	419	1.95	1	-3.62	1.84	-1.55	1.29	-11.00	408.00
108589	15	419	1.95	2	-4.01	1.23	-2.02	1.85	-12.78	406.22
108589	15	419	1.95	3	-5.02	1.06	0.76	3.26	-10.96	408.04
108589	15	419	1.95	4	-1.77	0.92	-0.70	3.47	-10.81	408.19
108589	15	419	1.95	5	-3.34	1.64	1.08	3.38	-9.45	409.55
148869	49	375	0.79	1	-3.62	0.37	-1.55	-0.23	-5.03	369.97
148869	49	375	0.79	2	-4.01	0.39	-2.02	-0.06	-5.94	369.06
148869	49	375	0.79	3	-5.02	0.10	0.76	1.25	-5.52	369.48
148869	49	375	0.79	4	-1.77	-0.52	-0.70	-1.10	-0.93	374.07
148869	49	375	0.79	5	-3.34	3.81	1.08	-0.31	-6.05	368.95

4.3. The analysis phase

After matching and adjusting the between-school matched control units, treatment effect estimation proceeded with the five multiply imputed data sets. For this empirical illustration, the analysis focused on the three research questions outlined above. Addressing these questions demonstrates how one can estimate the overall average treatment effect for the treated (ATT), school-level variance in the ATT, and explore factors related with school-level and student-level heterogeneity in the ATT. One should note that the estimation focused on the effect of assignment to algebra in 8th grade, which may not necessarily equate to a year of algebra course content exposure. As a result, the effect estimates are best thought of as intent-to-treat (ITT) estimates rather than treatment-on-the-treated (TOT) effects (Angrist, Imbens, & Rubin, 1996).

4.3.1. Does assignment to 8th grade algebra affect average student performance on the CAHSEE?

If one were to use the original sample of algebra and pre-algebra students and estimate the effect of 8th grade algebra with an unconditional two-level hierarchical linear model (HLM; see Equation 3.5), the resulting naïve average treatment effect estimate would be 38.87 scale score points, or roughly one standard deviation. Furthermore, the estimate of between-school effect variance would be 140.18 ($\sqrt{\tau}$ = 11.84) scale score points. This naïve estimate is riddled with selection bias. Using the same unconditional two-level HLM with the matched and adjusted data resulted in an overall ATT estimate of 9.00 scale score points (se = 0.86) and a between-school effect variance estimate of 20.64 ($\sqrt{\tau}$ = 4.54) scale score points. Since the matching diagnostics indicated some remaining pre-treatment covariate group differences after matching, the unconditional model estimates may also be biased. To account for remaining covariate

imbalance, I included the group-mean centered student pre-treatment covariates (**X**) used in the propensity score model in the following conditional model to estimate average treatment effects:

$$\begin{aligned}
&\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + \boldsymbol{\beta_2} \mathbf{X_{ij}^{gp}} + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2), \\
&\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_0), \\
&\qquad\qquad \beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_1), \\
&\qquad\qquad \boldsymbol{\beta_2} = \boldsymbol{\gamma_{20}}.
\end{aligned}$$

With the conditional model, the overall ATT estimate (γ_{10}) was 7.33 scale score points (se = 0.78) and the between-school effect variance estimate (τ_1) was 19.96 ($\sqrt{\tau_1}$ = 4.47) scale score points. The small difference between the unconditional model estimates and the conditional model estimates suggests that, for the most part, the matching did a good job removing pre-treatment covariate bias. Based on the conditional model estimates, the average effect of assignment to 8th grade algebra was a statistically significant positive effect size of about 0.23 standard deviations.

4.3.2. Does the effect of assignment to 8th grade algebra differ across schools?

The between-school effect variance estimate from the conditional model suggests meaningful school-level heterogeneity in the ATT. While the overall ATT was 7.33 scale score points, the ATT in a school with an ATT one standard deviation larger than the average school would be 11.80 scale score points. Conversely, the ATT in a school with an ATT one standard deviation smaller than the average school would only be 2.86 scale score points. Heterogeneity in school-level ATT was apparent when looking at each school's empirical Bayes estimated ATT (see Figure 4.12). A positive statistically significant ATT was estimated for 35 of the 50 schools, with the estimated ATT significantly below the overall ATT in nine schools and significantly above the overall ATT in four schools.

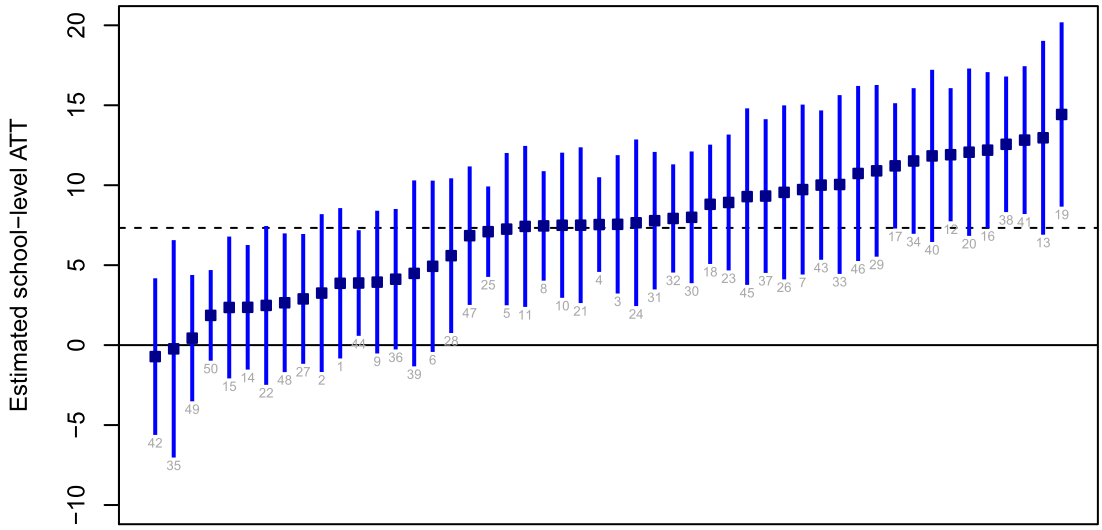

Figure 4.12. Empirical Bayes estimated ATT and approximate 95% confidence interval for each school.

4.3.3. Are certain factors associated with heterogeneity in the average effect of assignment to 8th grade algebra?

Given meaningful between-school heterogeneity in the ATT, the interest turned to exploring factors associated with this heterogeneity. It is important to note that the two-stage matching approach was designed to estimate causal effects at the site level, but efforts to understand heterogeneity in the causal effects falls outside the approach's causal purview into an exploratory or descriptive analysis. For the empirical demonstration I explored three sources of potential heterogeneity: (1) effect heterogeneity across student-level subgroups; (2) effect heterogeneity across school characteristics; and (3) effect heterogeneity associated with classroom composition measured at both the student level and the school level. If student-level subgroup heterogeneity existed, schools that disproportionately placed "high benefit" subgroups

in algebra would have a higher ATT, everything else equal. Similarly, if certain school characteristics were related to the magnitude of the ATT then schools with those characteristics would have a higher ATT. For example, higher achieving schools may have higher academic expectations and these expectations may improve the effectiveness of algebra instruction. I explicitly explored the role of classroom composition to examine possible SUTVA breakdowns due to peer effects and the concern expressed by some researchers (Loveless, 2008) that expanding algebra to unprepared students will water-down instructional quality for more prepared students.

One appealing way to explore effect heterogeneity in a multisite design with a two-level HLM was described in Bloom, Hill & Riccio (2003), where treatment-by-unit subgroup interactions were included at level-1 and treatment-by-site characteristic interactions were included at level-2. The formal model has the general following form:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j} D_{ij} + \boldsymbol{\beta}_{2j} \mathbf{X}_{ij}^{gd} + \boldsymbol{\beta}_{3j} D_{ij} \mathbf{X}_{ij}^{gd} + e_{ij},$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \boldsymbol{\gamma}_{01} \mathbf{S}_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10} + \boldsymbol{\gamma}_{11} \mathbf{S}_j + u_{1j},$$

$$\boldsymbol{\beta}_{\mathbf{2j}} = \boldsymbol{\gamma}_{\mathbf{20}},$$

$$\boldsymbol{\beta}_{\mathbf{3j}} = \boldsymbol{\gamma}_{\mathbf{30}},$$

where $\boldsymbol{\gamma}_{30}$ captures the vector of differential treatment effects across student subgroups, assumed fixed across schools for ease of estimation, and $\boldsymbol{\gamma}_{11}$ captures the vector of differential treatment effects across school characteristics. Grand-mean centering the level-1 covariates facilitates exploration of site-level effect heterogeneity, controlling for differential treatment effects at level-1 (Bloom et al., 2003).

I included a host of student subgroup indicators in the model, including subgroups that represent standard demographic groups and subgroups based on student performance on the 7th

grade mathematics CST. The latter subgroup was of particular interest given the concern that algebra may not be appropriate for students who have not mastered earlier mathematics content. I included two school characteristics in the model: the school's academic achievement index used to create site clusters and a proxy of the school's algebra placement selectivity. I constructed the selectivity proxy by calculating the observed proportion of "average performing" students in each school who were in 8th grade algebra, where "average performing" was defined as students who scored in the basic proficiency level on the 7th grade mathematics CST and received a C in their 7th grade mathematics course. For the analysis, the school's observed proportion of algebra placement was converted to log-odds for more normal distribution properties. Note that higher values correspond to less selective, or more egalitarian, algebra placement practices and lower values correspond to more selective, or restrictive, placement.

The potential role of classroom composition was explored in three ways. At the student-level, I classified students into one of three categories based on their 7th grade mathematics CST performance relative to their 8th grade mathematics classroom peers. Students were classified as below their classroom peers if their CST scale score was 25 points (approximately half a standard deviation) below their classroom mean score, or above their classroom peers if their score was 25 points above their classroom mean score. Students with 7th grade scores within 25 points of their classroom mean were classified as average students. If teachers adapted classroom instruction to match the average or median student, this relative rank classification allows one to examine whether the effect of algebra was different for students who found the instruction/content too advanced (the below average students) or too remedial (the above average students). At the school level, I examined two measures of average algebra classroom composition: the school's mean 7th grade mathematics CST score for algebra classrooms and the

school's mean heterogeneity in 7th grade mathematics CST performance within algebra classrooms. The algebra classroom heterogeneity indicator was based on the 7th grade math CST interquartile range within each algebra classroom. These two school-level measures test whether the average effect of algebra was related to average overall peer group prior achievement levels and/or the degree to which peer groups are comprised of students with similar or different prior achievement levels.

I estimated the interaction effects in stages to monitor whether the addition of different treatment interactions altered the estimation of other interactions. Specifically, I estimated four models, where the first model only includes the student-level subgroup interactions. Building off the first model, I added school-level characteristic interactions for the second model, and added the student-level relative classroom standing measure in the third model. For the fourth model, I added the school-level classroom composition measures. Results from these model are presented in Table 4.8.

Among the treatment-by-student subgroup interactions, statistically significant treatment effect heterogeneity was limited to two subgroups: 7th grade math performance levels and race/ethnicity. Across the different model specifications, students who scored proficient or above on the 7th grade math CST benefited more from taking algebra versus pre-algebra. On average, the algebra effect for proficient students was about 3-4 scale score points higher than students who scored basic on the 7th grade CST, everything else equal. Conversely, students who scored far below basic (FBB) on the 7th grade math CST did not benefit as much from taking algebra, on average. This limited algebra effect for low performing students was sizable and statistically significant once the differential effect of relative classroom standing was taken into account (models 3 and 4). These results are consistent with the argument that students who master 7th

grade mathematics content will benefit from algebra in 8th grade while those who struggle with 7th grade mathematics may not. Even after matching and controlling for differential effects across student subgroups, the effect of algebra was smaller for African American/black students. The differential effect was statistically significant and consistent across the four model specifications. Further research is needed to understand why 8th grade algebra might be less effective for African American/black students.

Independent of differential algebra effects across student subgroups, the effect of algebra depended on the student's relative classroom standing. The results from models 3 and 4 suggest that students who entered an 8th grade algebra classroom with 7th grade math achievement above the average student in the classroom did not benefit as much from algebra relative to a similar student in an algebra classroom with more students of equal mathematics achievement level. The algebra effect for students who entered a classroom with lower mathematics achievement than the classroom average was not statistically different from the effect for average students. This result is consistent with the argument that increasing access to algebra for students less proficient in 7th grade mathematics content may water down instructional quality for more advanced students. It is important to note, however, that while some differential effects existed at the student level, on average, even student subgroups that experienced smaller algebra effects were likely to perform better on the CAHSEE if in algebra instead of pre-algebra (i.e., the average effect estimates were still positive). Figure 4.13 plots the expected algebra effect and approximate 95% confidence interval for each of the subgroups based on model 4. For all but one subgroup, FBB CST students, the confidence interval lower bound was positive.

Table 4.8. Differential treatment effect estimates based on four model specifications.

	Model 1		Model 2		Model 3		Model 4	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Grand-Mean ATT	6.83	(0.72) *	6.99	(0.73) *	9.73	(0.85) *	9.74	(0.87) *
Treatment-by-Student Subgroup Interactions								
7th Grade CST PL: FBB	-2.90	(1.95)	-2.91	(1.96)	-8.80	(2.34) *	-8.82	(2.35) *
7th Grade CST PL: BB	0.16	(1.12)	0.16	(1.13)	-4.04	(1.30) *	-4.05	(1.31) *
7th Grade CST PL: PP	3.24	(1.10) *	3.08	(1.13) *	4.03	(1.23) *	3.94	(1.23) *
Female	1.16	(0.82)	1.12	(0.82)	0.76	(0.81)	0.77	(0.81)
Afr. Am./Black	-5.02	(2.12) *	-5.07	(2.16) *	-4.94	(2.15) *	-4.91	(2.15) *
Hispanic/Latino	-2.49	(1.70)	-2.37	(1.75)	-2.24	(1.74)	-2.23	(1.74)
Other Race/Ethnicity	-2.46	(2.12)	-2.50	(2.12)	-2.54	(2.11)	-2.46	(2.12)
English Learner	-0.87	(1.70)	-0.88	(1.71)	-0.52	(1.71)	-0.62	(1.71)
Initial-Fluent EP	0.25	(1.77)	0.18	(1.77)	0.26	(1.76)	0.25	(1.76)
Reclass.-Fluent EP	-1.04	(1.40)	-1.07	(1.41)	-1.03	(1.41)	-1.08	(1.41)
In Gifted/Talented Program	0.85	(1.19)	0.82	(1.21)	0.53	(1.20)	0.53	(1.20)
Ever Suspended	0.54	(1.50)	0.60	(1.50)	0.56	(1.50)	0.54	(1.50)
Student w/Disabilities	3.90	(3.04)	4.01	(3.04)	4.65	(3.03)	4.64	(3.03)
Free/Reduced Meals Eligible	-1.65	(1.06)	-1.57	(1.06)	-1.23	(1.06)	-1.19	(1.06)
7th Grade Attendance Rate	0.48	(0.42)	0.48	(0.42)	0.50	(0.42)	0.49	(0.42)
Treatment-by-School Characteristics Interactions								
School Achievement Index	--		0.27	(0.92)	-0.03	(0.94)	-0.28	(1.52)
School Algebra Selectivity	--		-0.92	(0.52)	-1.15	(0.53) *	-0.91	(0.83)
Treatment-by-Student Relative Classroom Standing								
Below Average Classroom Peers	--		--		3.11	(2.01)	3.12	(2.01)
Above Average Classroom Peers	--		--		-3.29	(1.25) *	-3.21	(1.25) *
Treatment-by-School Average Classroom Composition Measures								
Average Classroom Achievement	--		--		--		0.03	(0.07)
Average Classroom Heterogeneity	--		--		--		-0.14	(0.11)

Note: the grand-mean ATT represents the conditional average algebra effect for the omitted reference groups. Those groups, depending on the model, include: students who scored in the 7th grade math CST "basic" performance level; males; whites; English only; non-GATE; never suspended in 7th grade; non-student w/ disability; not eligible for a meals program; at classroom peer average 7th grade performance.
* point estimate is more than twice the standard error.

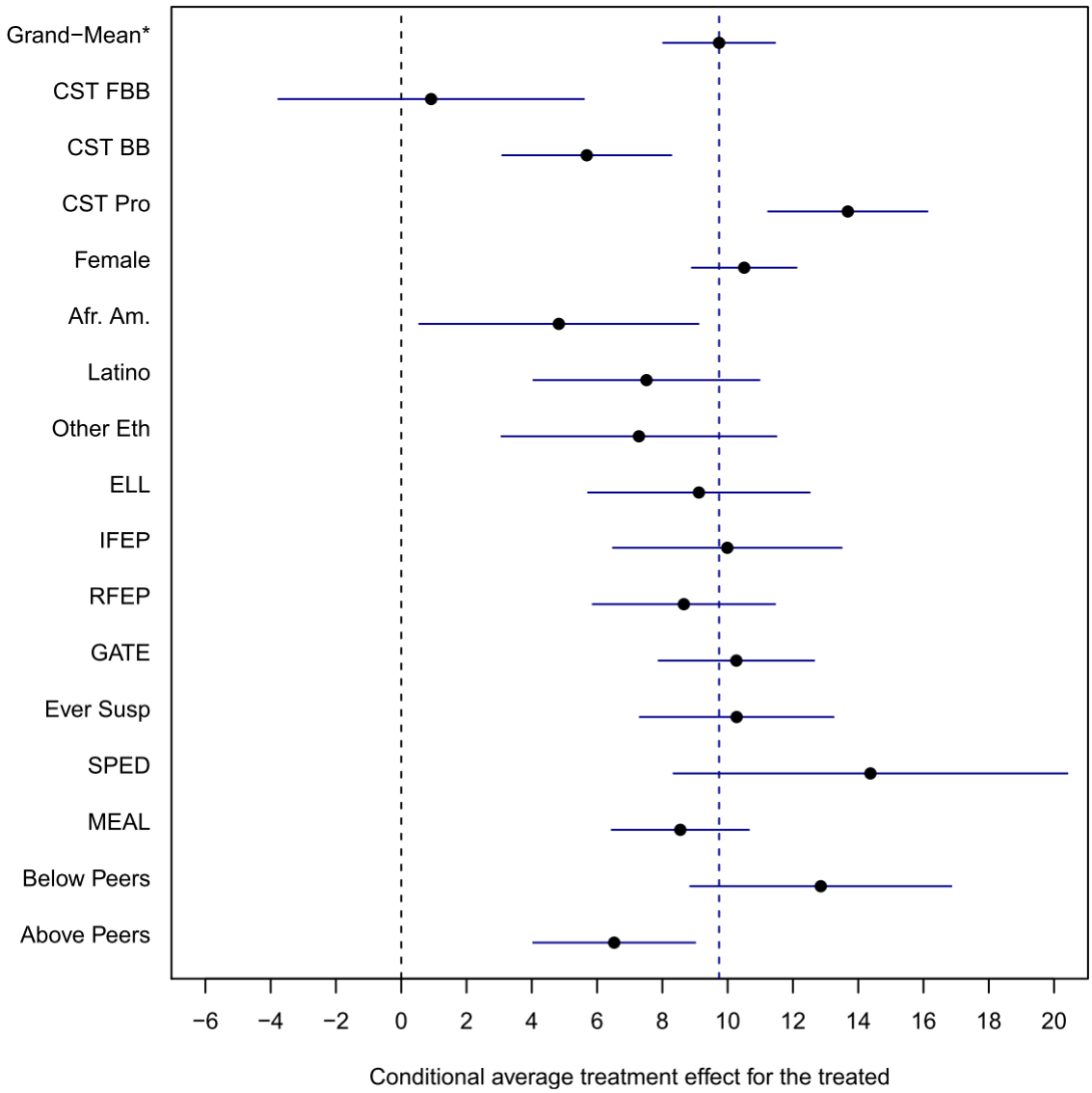

Figure 4.13. Estimated conditional average treatment effects and approximate 95% confidence intervals for select student subgroups. Displayed estimates based on model 4 results.

* grand-mean reflects the conditional main effect of algebra assignment for the reference groups.

At the school level, neither the school characteristics nor the algebra classroom composition measures had a strong association with the ATT. The estimated relationship between the algebra selectivity indicator and school-level average effect was consistently

negative across models 2-3, but only statistically significant in model 3. While a negative relationship may exist—meaning schools with less restrictive algebra placement practices had a lower average treatment effect—the relationship was either clouded by the inclusion of algebra classroom composition measures or the data did not have enough power to detect a statistically significant association. Somewhat surprisingly, a school's average treatment effect was not significantly associated with the school's algebra classroom composition measures. If true, these null results suggest that, while peer effects at the student level (as measured by a student's relative standing) may have influenced the effectiveness of 8th grade algebra, average peer composition did not help explain variation in the school-level ATT. Including the student subgroup interactions accounted for about a third of the between-school ATT variance, and adding the school-level interactions into the model did not alter the amount of remaining between-school variance.

4.4. Comparison of results across different estimation methods

The previous sections in this chapter described the process by which I used the two-stage matching strategy to estimate the effects of assignment to 8th grade algebra. In this section, I compare the overall results from the two-stage matching strategy to alternative estimation methods. First, I focus on alternative specifications within the two-stage matching framework. Second, I compare the results to more standard estimation methods.

4.4.1. Estimated effects under alternative specifications within the two-stage matching framework

To examine sensitivity of the overall ATT to different specifications within the three phases of the two-stage matching strategy, I re-ran the analysis with the following changes:

- Use a single-level logistic regression model to estimate the propensity score;
- Use a random intercept (RI) two-level logistic regression model to estimate the propensity score;
- Skip the adjustment phase;
- Only impute one adjustment for the between-school matched control students; and
- Restrict the analysis to within-school matches.

The resulting overall average treatment effect and between-school effect variance estimates based on the different specifications are presented in Table 4.9. For each alternative specification, two different sets of estimates are presented: one set based on a two-level HLM that included the treatment indicator but no other covariates at level-1 (unconditional model) and another set based on a two-level HLM that included the treatment indicator and the other key pre-treatment covariates (conditional model).

In general, the overall ATT estimate and between-school ATT variance estimate was stable across alternative specifications to the two-stage matching framework. While the full two-stage matching strategy utilized a random-intercept-and-slope (RIS) propensity score model in the design phase, effect estimates based on a more straightforward—and computationally simpler—single-level or RI two-level model were not appreciably different from the full two-stage matching design. For example, the overall ATT estimates from a conditional model were 7.33 for the full design, 7.20 with a single-level propensity model, and 6.84 with a RI model.

Similarly, the between-school ATT variance estimates were, respectively, 19.96, 19.35, and 19.15. It is also of interest to note that changes in the propensity score model did not result in meaningful changes in the number of students matched using the two-stage strategy. For each model specification, the matched sample size remained around 10,500 students.

Table 4.9. Average treatment effect and between-school variance estimates based on alternative specifications to the two-stage matching framework.

	Overall ATT				School-Level ATT	
	N	Estimate	SE	T-Value	Variance	Std. Dev
2-Stage Matching						
Unconditional Model	10,538	9.00	0.86	10.41	20.64	4.54
Conditional Model	10,538	7.33	0.78	9.41	19.96	4.47
2-Stage Matching using a Single-Level Propensity Score Model						
Unconditional Model	10,500	8.82	0.85	10.38	19.67	4.44
Conditional Model	10,500	7.20	0.77	9.38	19.35	4.40
2-Stage Matching using a RI Two-Level Propensity Score						
Unconditional Model	10,576	8.44	0.85	9.96	18.23	4.27
Conditional Model	10,576	6.84	0.78	8.76	19.15	4.38
2-Stage Matching with No Adjustment						
Unconditional Model	10,538	7.53	0.88	8.59	22.32	4.72
Conditional Model	10,538	5.79	0.79	7.30	22.19	4.71
2-Stage Matching with 1 Imputed Adjustment						
Unconditional Model	10,538	8.82	0.84	10.51	19.74	4.44
Conditional Model	10,538	7.14	0.75	9.48	19.23	4.38
2-Stage Matching with Analysis Restricted to Within-School Matches						
Unconditional Model	6,024	9.21	0.86	10.69	16.29	4.04
Conditional Model	6,024	6.40	0.77	8.29	16.55	4.07

Note: RI = random-intercept.

Skipping the adjustment phase resulted in slightly smaller average effect estimates (5.79 vs. 7.33 for the conditional model) and slightly larger between-school variance estimates (22.19

vs. 19.96 for the conditional model). This suggests that the between-school matching resulted in treatment units matched to control units in higher achieving schools, on average. So if those "inflated" control unit outcomes are not adjusted downward, the estimated average treatment effect is lower. Changing the number of adjustments from five to one, did not result in a meaningful difference in the effect estimates. Since multiply imputing the adjustment allows one to incorporate uncertainty in the school effect estimates into the analysis, however, the single adjustment estimation had slightly smaller standard errors than the full analysis with five imputed adjustments.

Small differences between the full two-stage estimates and estimates that ignored the adjustment phase raise questions about the between-school matches. Therefore, I also estimated the overall ATT and the between-school variance based on just the within-school matched students. An analysis based solely on within-school matches should be devoid of school-level bias, but limiting the sample to acceptable within-school matches could under-estimate the ATT given student-level effect heterogeneity. Looking back at Tables 4.4 and 4.5 also indicates that student-level covariate balance was a little worse for the within-school matches relative to the full matched sample. The unconditional model effect estimates based on the within-school matches may reflect this covariate imbalance, with the overall ATT estimate of 9.21 slightly higher than the two-stage matching estimate of 9.00 scale score points. When residual covariate imbalance is adjusted for with the conditional model, the within-school estimate was lower (6.40) than the full matching estimate (7.33), which could reflect the differential effectiveness of 8th grade algebra for the students who are part of the within-school matches (typically lower performing) and those who are part of the between-school matches (typically higher performing). Restricting the analysis to within-school matches also resulted in less between-school effect

variance compared to the two-stage matching. This variance difference could reflect unobserved school-level bias in the two-stage matching analysis or the restricted, more homogeneous, within-school sample.

Another way to investigate school-level effect variance is to compare the school-specific ATT estimates across the estimation methods. Scatterplots of the school-level ATT estimates based on the full two-stage matching strategy and the alternative specifications are presented in Figure 4.14. Points below the 45-degree line indicate schools where the two-stage matching estimate was larger than the alternative estimate, while points above the 45-degree line indicate the opposite. No particular trends or outliers are apparent, suggesting that effect estimate differences from the alternative specifications did not have unique implications for specific schools. One exception was the ATT estimate for school 4 when the adjustment phase was skipped. For the vast majority of schools, skipping the adjustment phase resulted in a lower ATT estimate relative to the full two-stage matching strategy. For school 4, however, skipping the adjustment phase resulted in a higher ATT estimate (9.47 vs. 7.54). School 4 did not have an abnormally high percentage of between-school matches, but was a high achieving school relative to other schools within site cluster 1 (see Figure 4.9). It is likely, therefore, that the between-school matches for school 4 algebra students drew from relatively lower performing schools and thus ignoring these school differences inflated the ATT estimate for school 4.

An examination of school-specific ATT estimates based on the within-school matching analysis versus the full two-stage matching analysis (See Figure 4.15) also reflects the overall comparisons from Table 4.9. For most schools, the within-school analysis produced a smaller average effect estimate relative to two-stage matching, which may reflect the restricted sample produced by within-school matching and differential impact of 8th grade algebra across students.

Schools that had within-school effect estimates farthest from their two-stage matching estimate (i.e., the schools farthest from the 45-degree line) were, generally, the schools with a relatively low proportion of within-school matches. These schools fell both above (e.g., school 35) and below (e.g., schools 32 and 39) the 45-degree line, however.

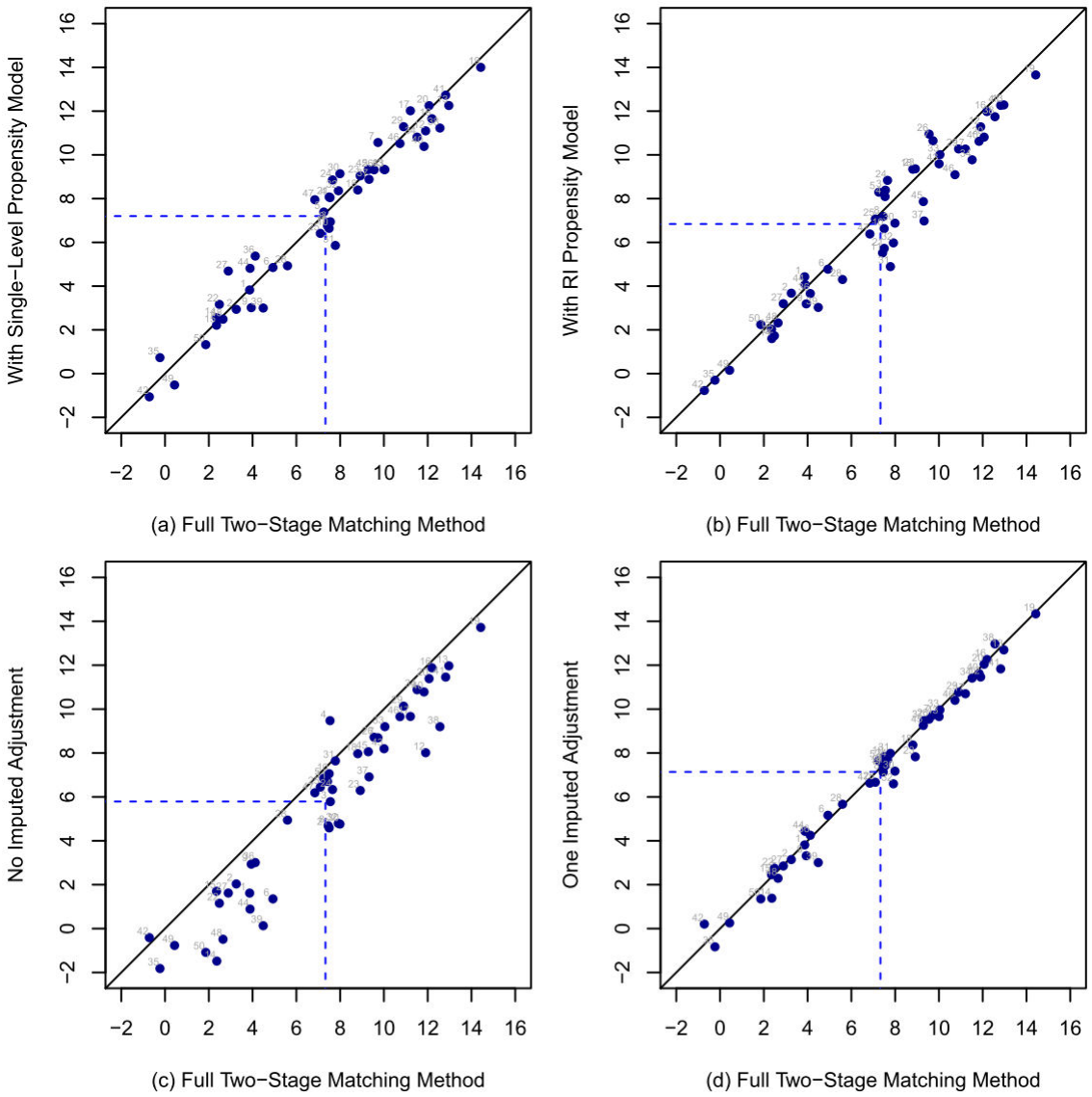

Figure 4.14. Comparisons of school-specific ATT estimates based on the two-stage matching strategy and alternative specifications.

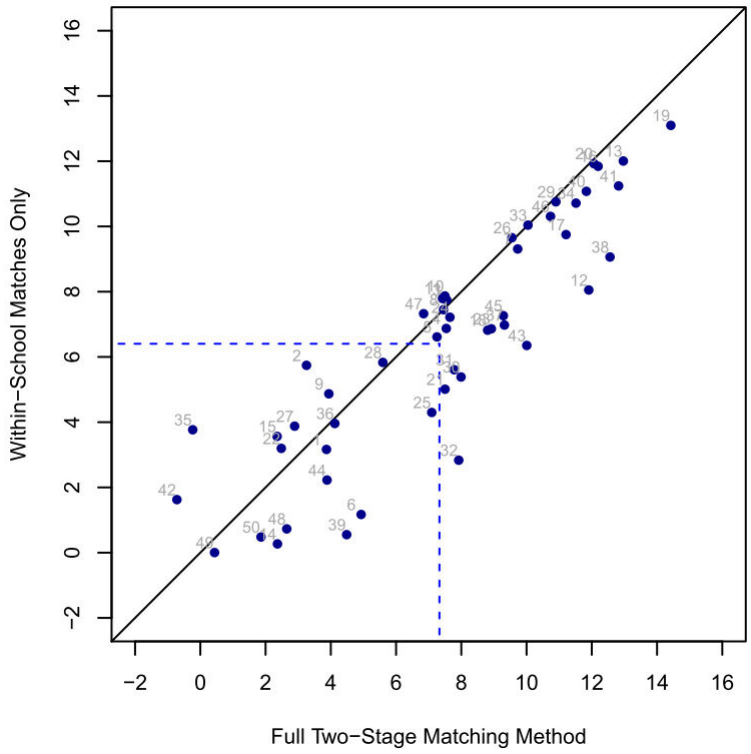

Figure 4.15. Comparisons of school-specific ATT estimates based on the two-stage matching strategy and within-school matching only.

4.4.2. Estimated effects under more standard estimation methods.

To compare effect estimates from the two-stage matching strategy to more traditional estimation approaches, I re-analyzed the data using a two-level HLM outcome model based on the following preprocessing in the design phase:

- No matching (i.e., the full sample);
- Pooled matching using a single-level propensity score model;
- Pooled matching using a random-intercept (RI) two-level propensity score model;
- Pooled matching using a random-intercept-and-slope (RIS) two-level propensity score model.

In all cases I skipped the adjustment phase, except for the full two-stage matching method. A two-level HLM outcome model based on the full sample is the simplest estimation method that still acknowledges the nesting of students within schools and a desire to estimate school-level average effects. Estimates based on pooled matching—matching that allows matches from any school without prioritizing matches from either within- or between-schools—is the simplest matching method that still acknowledges a desire to equate treatment and control student pre-treatment covariates in a preprocessing phase. Differences in the propensity score model represent different levels of propensity score estimation complexity/flexibility, as well as different degrees of concern regarding school-level heterogeneity in the assignment mechanism.

The resulting overall average treatment effect and between-school effect variance estimates based on the different estimation methods are presented in Table 4.10. For each method, two different sets of estimates are presented: one set based on a two-level HLM that includes the treatment indicator but no other covariates at level-1 (unconditional model) and another set based on a two-level HLM that includes the treatment indicator and the other key pre-treatment covariates (conditional model). One should note that the two-level model estimates without matching are what I referred to earlier in this chapter as the naïve model estimates. Both the overall ATT estimate and the between-school effect variance estimate from the naïve model were grossly inflated due to selection bias.

Table 4.10. Average treatment effect and between-school variance estimates based on different estimation methods.

	Overall ATT				School-Level ATT	
	N	Estimate	SE	T-Value	Variance	Std. Dev
2-Stage Matching						
Unconditional Model	10,538	9.00	0.86	10.41	20.64	4.54
Conditional Model	10,538	7.33	0.78	9.41	19.96	4.47
Two-Level Model without Matching						
Unconditional Model	19,063	38.87	1.74	22.30	140.18	11.84
Conditional Model	19,063	5.36	0.69	7.82	14.17	3.76
Pooled Matching with a Single-Level Propensity Score Model						
Unconditional Model	8,920	15.16	1.28	11.89	58.05	7.62
Conditional Model	8,920	6.23	0.80	7.82	19.20	4.38
Pooled Matching with a Two-Level RI Propensity Score Model						
Unconditional Model	7,314	10.18	1.08	9.42	40.09	6.33
Conditional Model	7,314	5.69	0.69	8.20	12.86	3.59
Pooled Matching with a Two-Level RIS Propensity Score Model						
Unconditional Model	7,348	11.54	0.80	14.49	14.69	3.83
Conditional Model	7,348	6.46	0.72	8.92	15.32	3.91

Notes: RI = random-intercept; RIS = random-intercept-and-slope.

Using a HLM to adjust for differences in the observed pre-treatment covariates (the conditional model) resulted in an overall ATT estimate (5.36) and between-school variance estimate (14.17) more aligned with the two-stage matching method. The effect estimates from the unmatched conditional model were, however, lower than from the other methods. Two factors may have contributed to this difference. First, not preprocessing the data via matching means covariate adjustment is highly dependent on the HLM parameter estimates and functional form. The estimated model may have resulted in over-correcting for covariate imbalance. Second, given effect heterogeneity at the student level, estimating effects from a model that uses

the full sample is not necessarily estimating the ATT, but a conditional-variance-weighted estimate (Morgan & Winship, 2007) that may not reflect the ATT.

Estimates based on a matched sample using pooled matching also deviated slightly from the two-stage matching estimates. Regardless of the propensity score model employed, the unconditional model estimate for the overall average ATT was larger than the two-stage matching estimate. Assuming the true overall ATT is equal to or below 9.00 scale score points, the two-stage matching approach appears to have done a slightly better job reducing pre-treatment covariate bias than the pooled matching options. When a conditional model was employed to adjust for covariate differences, the overall ATT estimates were, like the unmatched estimates, a little below the two-stage matching estimate. The estimates based on pooled matching with a RI propensity score model stood out as particularly low, with an overall ATT estimate of 5.69 scale score points and a between-school effect variance estimate of 12.86. The restricted sample that resulted from the pooled matching was one possible explanation for the low effect estimates. The pooled matching based on the RI propensity score model, for example, only retained 3,657 algebra students, while the two-stage matching method retained 5,269 algebra students.

A comparison of site-specific ATT estimates indicates similar small differences in the estimation methods (see Figure 4.16), on average. For some schools the estimates differed noticeably, however. For example, school 35 consistently had a lower ATT estimate with the two-stage matching method compared to the other estimation methods. School 35 had a relatively high proportion of students in algebra and under the two-stage matching method very few treatment students were matched to control students within the same school. The school's ATT estimate based on within-school matching (see Figure 4.15) was also higher than the

estimate from the two-stage matching method. School 32, on the other hand, consistently had a higher ATT estimate with the two-stage matching method compared to the other estimation methods and the within-site matching estimate. Since the true ATT for these schools is not known, it is not clear whether the two-stage matching process resulted in a less or more biased ATT estimate compared to the other methods. On average, at least, the effect estimates were fairly stable across methods.

Given an interest in exploring effect heterogeneity, I also compared estimates of differential treatment effects based on the two-stage matching approach (reported in Table 4.8) to alternative approaches: no matching, pooled matching, or within-school matching. In general, the subgroup effects are similar, or at least in the same direction, across the different matching approaches. The identification of some subgroup effect estimates as statistically significant depended on the matching method used. Since the two-stage matching method retained more units in the analysis than the pooled or within-school matching, standard errors were consistently smaller for the two-stage matching estimates. One notable difference in the two-stage matching estimates was the statistically significant positive effect for high achieving students (those who scored proficient or above on the 7th grade mathematics CST). Compared to the alternative methods, the estimated effect for this subgroup was larger and statistically significant with the two-stage matching method. This could reflect the fact that compared to pooled and within-school matching, the two-stage matching approach retained more of the high achieving students in the analysis. Since the true effects are unknown, one cannot say whether the two-stage matching approach actually does a better job estimating effect heterogeneity, but this comparison does suggest that the matching method choice can have implications for conclusions about

sources of significant effect heterogeneity. At the very least, the two-stage method provides an alternative, and conceptually appealing, option for probing sources of heterogeneity.

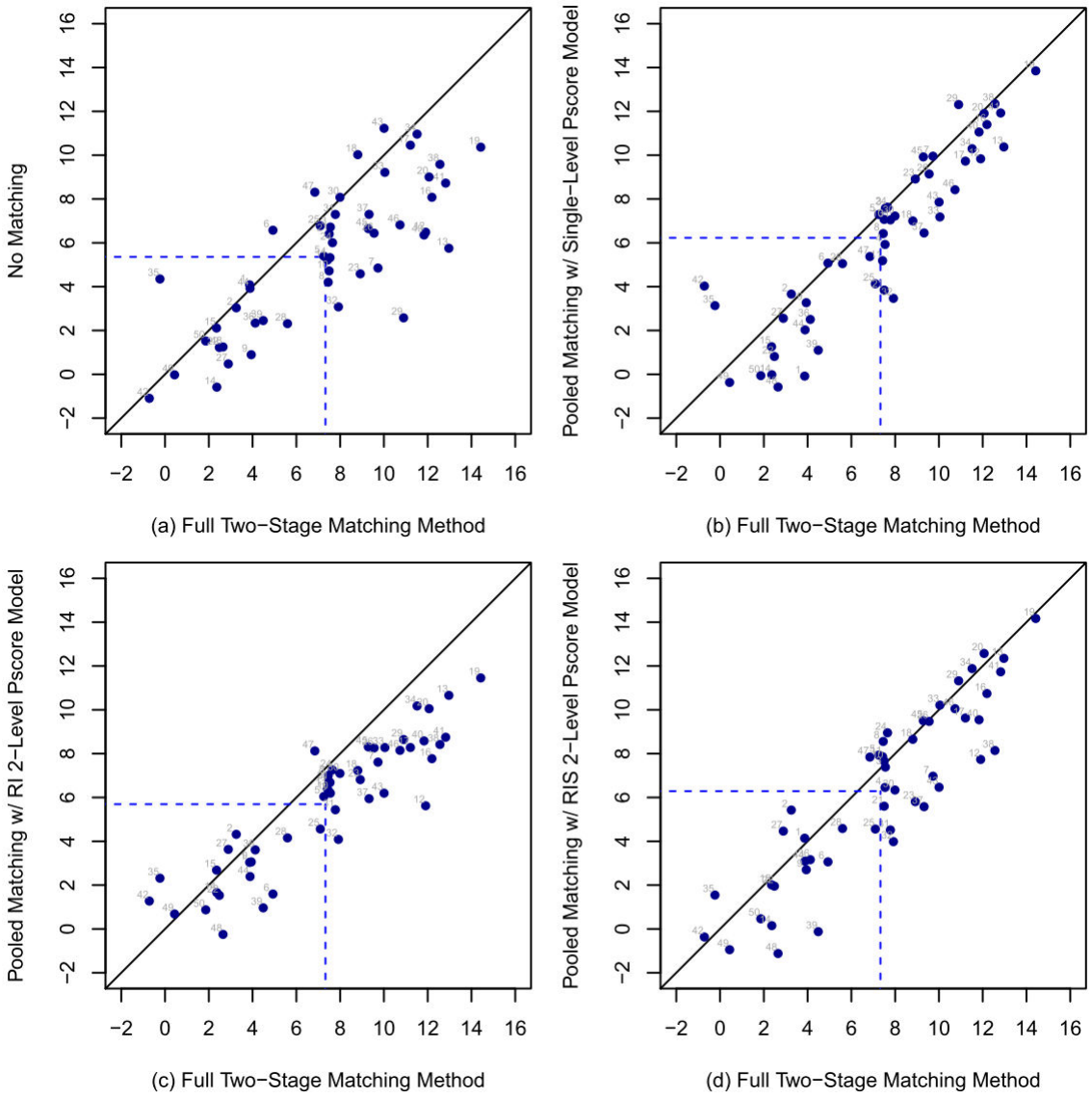

Figure 4.16. Comparisons of school-specific ATT estimates based on the two-stage matching strategy and different estimation methods.

4.5. Summary of key findings from the empirical illustration

In this chapter I illustrated how the proposed two-stage matching strategy can be applied to estimate the effect of assignment to 8th grade algebra. In this application, two-stage matching increased the percentage of matched treatment units from 57% based on within-site matching to 95% by including between-site matching as a secondary alternative. After matching, average pre-treatment covariate differences between the treatment and control groups were substantially reduced, although some residual covariate differences—particularly within some sites—warranted additional covariate adjustment in the analysis phase.

Analysis of the matched data indicated that, on average, assignment to algebra in 8th grade instead of a pre-algebra course increased student performance on the mathematics high school exit exam by a little over seven scale score points, corresponding to a fifth of a standard deviation. Significant effect heterogeneity existed at both the student and school levels, however. School-specific average effect estimates ranged from zero to over ten scale score points. Across students, students who demonstrated proficiency of 7th grade mathematics content experienced larger effects from algebra than students who struggled with 7th grade mathematics content. Additionally, peer effects may have mediated the effect of algebra by dampening the effect of algebra for students with higher mathematics understanding relative to their classroom peers. This interplay between relative classroom standing and differential treatment effects also emerged in research on the use of double-dose algebra in Chicago (Nomi & Allensworth, 2009). Accounting for effect heterogeneity at the student level accounted for about a third of the between-school effect variance, but school-level indicators of school context and classroom composition were not significantly associated with the effect of 8th grade algebra. The importance of effect heterogeneity at level-1 for explaining site-level effect heterogeneity was

also found in the Bloom et al. (2003) study of welfare-to-work programs. Future research on this topic should seek to examine what factors, such as instructional content, quality, and supports account for the observed variation in effects across schools. More work is also needed to better understand why and how peer effects may mediate the effect of 8th grade algebra.

While this analysis found a positive average effect for assignment to 8th grade algebra, it is important to keep in mind that the results depend on the assumption of strongly ignorable treatment assignment. The estimated effects may be a manifestation of unobserved confounders. A sensitivity analysis, not presented here, indicated that if a moderately important unobserved confounder existed (e.g., an unobserved pre-treatment factor with independent correlations to treatment assignment and the outcome as strong as 7th grade math GPA) then the true average ATT would be a substantively meaningless size of about two to three scale score points. Additionally, analysis of a pseudo-nonequivalent outcome measure (ELA performance on the high school exit exam) suggests that unobserved confounders could account for about half of the estimated average algebra effect.

This demonstration also compared the estimated effects from the proposed two-stage matching strategy to alternative specifications and methods. Overall, the effect of 8th grade algebra was fairly robust to these alternative specifications and estimation methods as long as residual covariate group imbalance was adjusted for in the outcome model. In the design phase, using a single-level or RI two-level model rather than an RIS two-level model to estimate the propensity score did not result in substantively different effect estimates. Regarding the adjustment phase, the effect estimates were slightly lower when no adjustment for between-school matching was conducted. Using one versus five imputations for school effect adjustment did not result in substantively different average effects, but the multiple imputation approach

produced slightly larger standard errors that reflect uncertainty in the adjustment process. Comparing results from the two-stage matching to within-school matching effect estimation in the analysis phase reinforced possible limitations of within-school matching. In particular, analysis restricted to the within-school matched data produced a somewhat smaller overall average treatment effect estimate and smaller between-site effect variance estimate than the two-stage matching approach. This difference was likely due to the limited matched sample that resulted from the more restrictive within-site matching, particularly given differential effects across units. Small differences in the effect estimates were also observed when the effect was estimated without matching or using pooled matching. The matching method choice may have implications for conclusions about sources of significant effect heterogeneity, however.

Overall, the empirical illustration showed the logic and strengths of the two-stage matching approach, but also suggests the additional complications involved with the approach may not justify the small differences in average effect estimates. Unfortunately, since the true effects of 8th grade algebra are unknown, the strength of the two-stage matching approach relative to these other estimation methods is not immediately clear. In the following chapter I present results from a series of simulation studies, where the estimates from these different approaches can be benchmarked against known true values.

Chapter 5

Simulation Study Results

In the previous chapter I compared effect estimates from the proposed two-stage matching strategy to other options for effect estimation. Since the true effects of 8th grade algebra were unknown in the empirical illustration, those comparisons only provide limited information regarding the utility of the proposed method. In this chapter I present results from a series of simulation studies that formally examined the performance of the proposed method and sensitivity to alternative specifications in each of the three phases. As described in Chapter 3, I conducted a Monte Carlo simulation study for each phase, based on 100 independent data replications with known treatment assignment and treatment effects. The design phase and analysis phase simulation studies were based on the same 100 data replications, while the adjustment phase simulation study was based on a separate draw of 100 data replications.

For the simulations, data were generated based on a sample size of 50 sites, with an average of 200 units per site. Treatment assignment was determined using five different assignment mechanism conditions:

- Random assignment (RA);
- Selection on unit-level observables (L1OB);
- Selection on unit- and site-level observables (L2OB);
- Selection on unit-level observables and site-level observables and unobservables (L2UN); and
- Selection on unit- and site-level observables and unobservables (L1UN).

As a contextual reference, consider potential outcomes under a treatment or control condition ($Y(1)_{ij}$ and $Y(0)_{ij}$) on a mathematics standardized test score for students nested within schools, where the researcher has an observed measure of prior academic achievement for each student (X_{ij}), an observed measure of student socio-economic status (Z_{ij}), and an observed composite measure of each school's overall instructional resources (S_j). Also important in determining a student's test score outcome, however, are an unobserved student motivation factor (U_{ij}) and an unobserved school instructional quality factor (V_j). The extent to which the observed and/or unobserved factors confound treatment effect estimation depends on the assignment mechanism condition.

This chapter is organized based on the guiding research questions described in Chapter 3:

1. How do different specifications in the design phase of the proposed method influence covariate balance?
2. How do different specifications in the design phase of the proposed method influence inferences about treatment effects?
3. How do different specifications in the imputation phase of the proposed method influence inferences about treatment effects?
4. How do treatment effect estimates from the proposed method compare to estimates obtained from more common matching-based and regression-based methods?

5.1. How design phase specifications influence covariate balance

The objective in the design phase is to improve covariate balance between the treatment and control groups by preprocessing the data prior to analysis. Given propensity score matching, the two main decision points are how to estimate each unit's propensity score and the type of

matching to employ. Three propensity score model conditions and three matching conditions were examined in the simulation study. For each crossed condition, covariate balance was assessed based on the within-site absolute standardized bias (ASB) and variance ratio (VR) as defined in Equation 3.1.

The propensity score conditions were a single-level logistic regression model (SL PS), a two-level RI logistic regression model (RI PS), and a two-level RIS logistic regression model (RIS PS). With a SL PS, model estimates are based on the data pooled across sites, so units with the same covariate values will have the same estimated propensity score regardless of site membership. With a RI PS, the relationship between the covariates and treatment assignment is fixed across sites, but units with the same covariate values can have a uniformly higher or lower estimated propensity score depending on site membership. With the RIS PS, on the other hand, both the average propensity score and the relationship between covariates and treatment assignment are site-specific. This means two units with the same covariate values can have different estimated propensity scores depending on their site membership.

The matching conditions were pooled matching (PM), within-site matching (WM), and two-stage matching (2SM). With PM, treatment units can be matched to any control unit regardless of site membership, which may be preferred if treatment assignment and/or the outcomes are invariant to site membership. With WM, treatment units can only be matched to control units within the same site, which may be preferred if treatment assignment and/or outcomes depend on site membership. With 2SM, treatment units are first matched to control units within the same site and then matches are allowed between sites within the same site cluster, which may provide an efficient compromise between PM and WM.

Improvements in covariate balance relative to balance in the unmatched data were fairly stable across the propensity score and matching conditions. For the unit-level covariates, each version of matching marked a significant improvement in balance for the observed covariates (*X* & *Z*) but not unobserved covariate (*U*). The grand-mean within-site ASB, averaged across replications, for each unit-level covariate is presented in Figure 5.1 (L1OB & L2OB assignment mechanism conditions) and Figure 5.2 (L2UN & L1UN assignment mechanism conditions) by condition. More detailed summary statistics for the balance results are reported in Tables A.1 through A.5 in the appendix. I focus on the grand-mean within-site ASB in this section because it best represents the primary measure of covariate balance within the average study site and the VR was consistently stable across conditions both in terms of absolute and relative magnitude.

While not presented in the figures below, it is interesting to point out that even under random assignment some within-site covariate imbalance existed without matching. On average, treatment and control groups differed by about 0.12 of a standard deviation on each of the three unit-level covariates, with the maximum within-site ASB between 0.30 and 0.40 of a standard deviation. This highlights the fact that under finite samples even random assignment can result in covariate imbalance. When interpreting balance improvement from the different propensity score and matching conditions, one can use balance under random assignment as a benchmark for balance improvement.

Under an assignment mechanism based on unit-level observables (Figure 5.1, left column), all matching versions reduced the grand-mean within-site ASB for *X* and *Z* from over 0.25 to below 0.15. The ASB was consistently lower for within-site matching and two-stage matching compared to pooled matching, although the differences were not substantively large. Within-site and two-stage matching also outperformed pooled matching in terms of the

maximum site-level ASB (see Table A.2), although even under the best results covariate ASB was above 0.25 in at least one site, on average. While small differences existed across matching conditions, balance improvement was robust to the propensity score model specification.

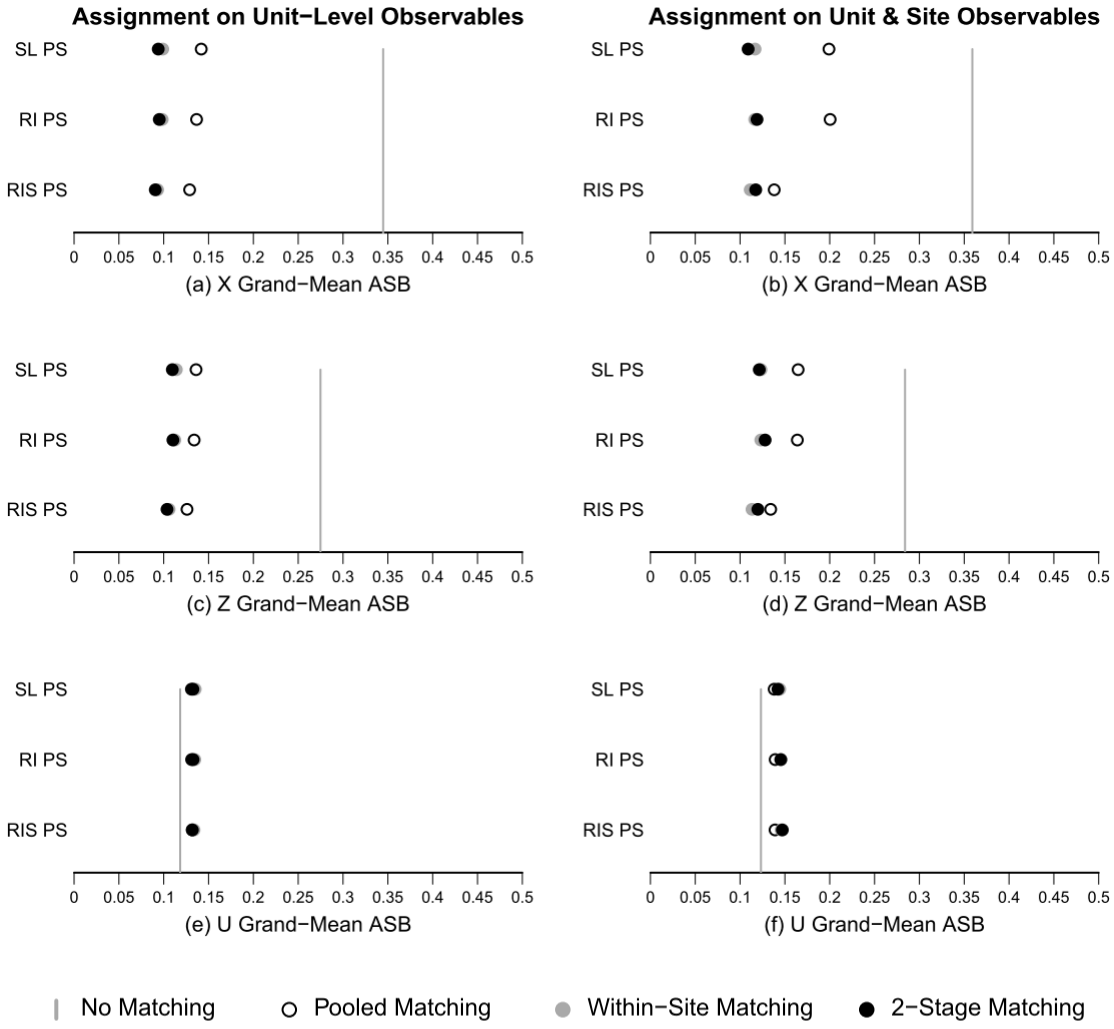

Figure 5.1. Grand-mean within-site ASB simulation results for unit-level covariates, by two assignment mechanisms with selection on observables

Similar results occurred when the assignment mechanism included selection on a site-level observable (Figure 5.1, right column) and selection on an unobservable (Figure 5.2), with

one notable exception. Balance improvement for *X* and *Z* was significantly better under the within-site and two-stage matching conditions compared to pooled matching, except when pooled matching was based on a RIS propensity score model. This suggests that when site-level confounders are present a RIS propensity score model may be the most appropriate option for improved covariate balance if pooled matching is the researcher's method of choice. Covariate balance based on within-site matching or two-stage matching, however, is less sensitive to the propensity score model decision. These simulation results also highlight the fact that none of the matching conditions will improve covariate balance for an unobserved unit-level confounder, and may even worsen balance (see Figure 5.2f).

For the site-level covariates, each version of matching marked a significant improvement in balance for the observed covariate (*S*) but performance regarding the unobserved covariate (*U*) differed across assignment mechanisms and the propensity score model employed. The grand-mean within-site ASB, averaged across replications, for each site-level covariate is presented in Figure 5.3 (L1OB & L2OB assignment mechanism conditions) and Figure 5.4 (L2UN & L1UN assignment mechanism conditions) by condition. More detailed summary statistics of the balance results are reported in Tables A.6 and A.7 in the appendix. When the assignment mechanism only depended on observed covariates (Figure 5.3), all matching methods significantly improve balance for the observed and unobserved covariate. Balance improvement was not sensitive to the propensity score model under within-site and two-stage matching, but balance improvement for the unobserved covariate under pooled matching was greatest with a single-level propensity score model. Since, the unobserved covariate was not part of the assignment mechanism, however, the mean ASB was below 0.10 of a standard deviation under all matching conditions.

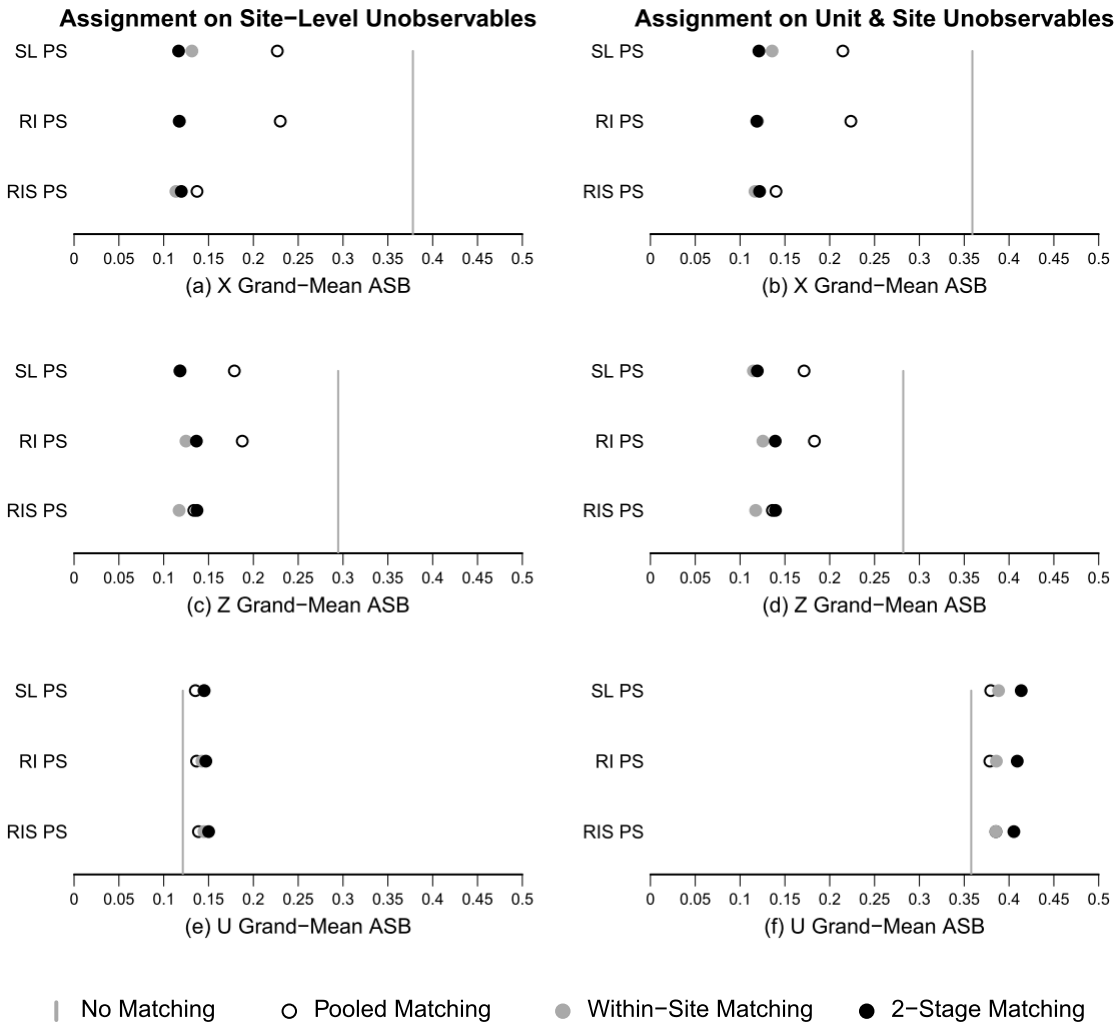

Figure 5.2. Grand-mean within-site ASB simulation results for unit-level covariates, by two assignment mechanisms with selection on unobservables

Balance performance for the unobserved site-level covariate really becomes important when it is part of the assignment mechanism (Figure 5.4). Under this type of assignment mechanism, the power of within-site matching was apparent. Within-site matching ensures that all observed and unobserved site-level covariates are balanced, and the simulation results reflected this defining property. In terms of the observed site-level covariate, ASB with two-stage matching was very similar to the within-site matching results and was consistently low

across the propensity score model conditions. Balance of the observed site-level covariate was also dramatically reduced with pooled matching, but to a slightly less dramatic degree compared to within-site and two-stage matching. Balance improvement under pooled matching was slightly better when a RI or RIS propensity score model was employed.

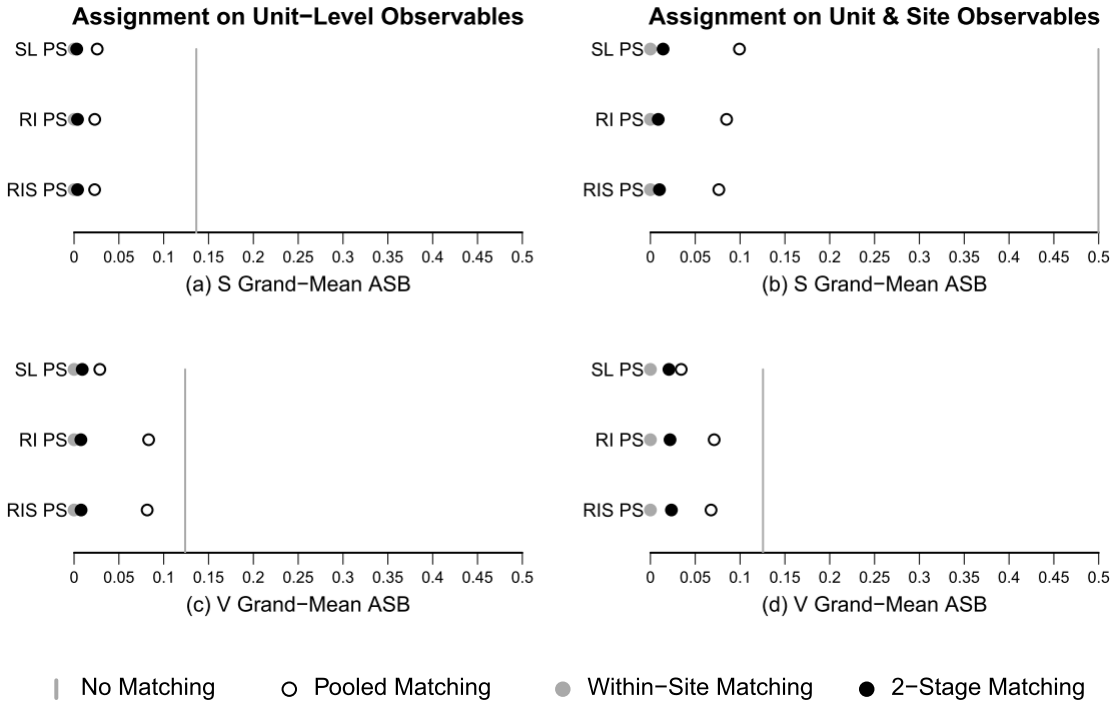

Figure 5.3. Mean ASB simulation results for site-level covariates, by two assignment mechanisms with selection on observables

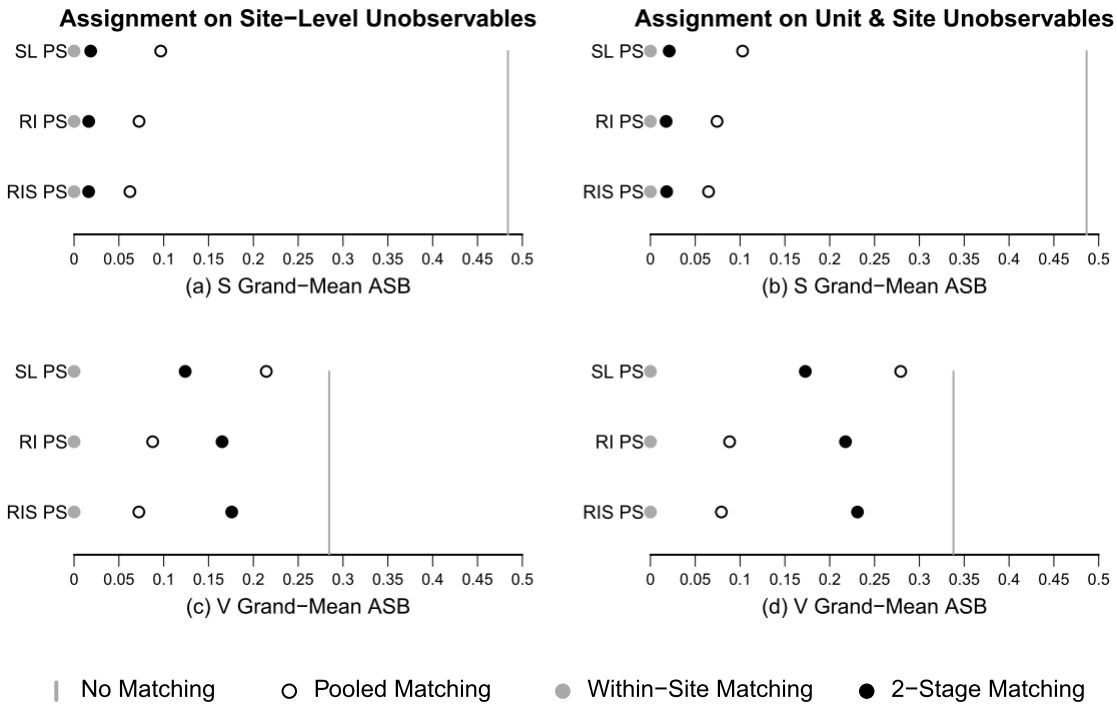

Figure 5.4. Mean ASB simulation results for site-level covariates, by two assignment mechanisms with selection on unobservables

While within-site matching eliminates imbalance in both the observed and unobserved site-level covariates, balance improvement for the unobserved covariate under two-stage and pooled matching depended on the propensity score model employed. With pooled matching based on a single-level propensity score model, substantive imbalance in the unobserved site-level covariate remained (ASB > 0.20). When pooled matching used a RI or RIS propensity score model, however, the ASB was below 0.10 of a standard deviation. Conversely, under two-stage matching, balance improvement for the unobserved site-level covariate was better when a single-level propensity score model was employed and poor with a RI or RIS model. This result was somewhat unexpected since the two-stage matching design was partially motivated to approximate within-site matching as closely as possible. The emphasis placed on balancing the observed unit-level covariates with the between-site matching stage, however, likely came at the

expense of limited balance improvement for unobserved site-level covariates. In fact, the purpose of the adjustment phase in the two-stage matching strategy is to account for any site-level bias inserted into the data by the between-site matching stage. This bias is at least partially represented by the unobserved site-level confounder.

While the two-stage matching method does not outperform within-site matching and only performs marginally better than pooled matching in terms of covariate balance, more treatment units are retained in the analysis with the two-stage method. Figure 5.5 displays the mean proportion of treatment units matched under each simulation condition. As intended, substantially more treatment units were matched using the two-stage matching method, regardless of assignment mechanism or propensity score model. For example, when treatment assignment was based on unit- and site-level observable covariates, just over 80% of the treatment units were matched within-site. Using the two-stage matching method increased the percent matched to over 95%. For the two-stage matching method, the match rate increased as the propensity score model became more flexible (i.e., went from fixed coefficients to RI to RIS). The opposite was true for the pooled matching condition.

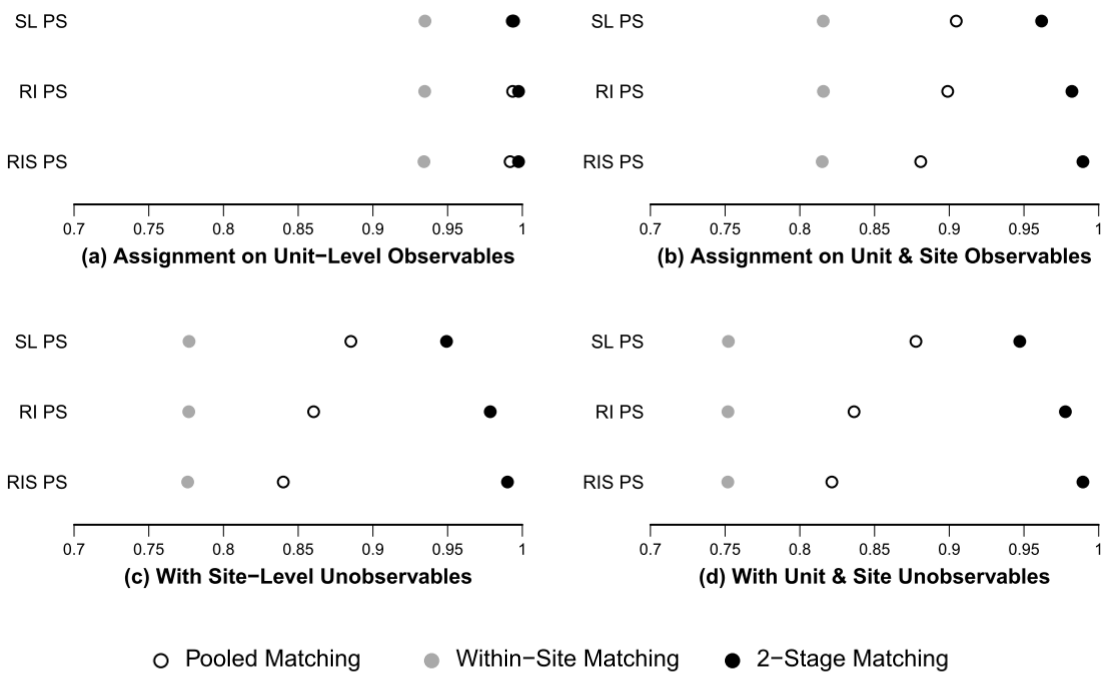

Figure 5.5. Mean proportion of treatment units matched across simulation replications, by assignment mechanism and design phase conditions

5.2. How design phase specifications influence treatment effect estimates

Results from the previous section identified certain conditions where the degree of covariate balance depended on the matching specification. Differences in covariate balance may not be large enough to influence the ultimate treatment effect estimation objective, however. In this section I examine the degree to which treatment effect estimates based on an unconditional two-level HLM differed across the design phase conditions. I focused on the estimation of the grand-mean average treatment effect for the treated (ATT) and site-level effect variance. To gauge relative performance under each condition, I examined mean bias from the true values and the root-mean-squared-error (RMSE). For ATT estimation, I also examined the coverage rate based on an approximate 95% confidence interval (± 2 × standard error). The full simulation results are presented in Table A.8 in the appendix. Note that the true grand-mean ATT value was

between 0.05 and 0.07, depending on the assignment mechanism (except the true ATT under the random assignment condition was zero), and both the average model-estimated standard error for the ATT and the Monte Carlo standard deviation was around 0.06 for all conditions.

Overall, all matching methods significantly reduced bias in the grand-mean ATT relative to no matching (i.e., relative to the naïve average treatment effect estimate). First, as expected, all conditions—including the naïve model—produced unbiased ATT estimates under random assignment, although the coverage rate was around 0.90 instead of the desired 0.95 rate. When assignment depended on observed covariates, ATT bias under all conditions was less than 0.05 (see Figure 5.6, top row), but performance of the two-stage matching method worsened with the more flexible propensity score models. For example, when assignment depended on the unit- and site-level observed covariates, ATT bias from the two-stage matching with a single-level propensity score model was 0.01, on par with bias from within-site matching. When two-stage matching used a RIS propensity score model, ATT bias was 0.05, while within-site matching and pooled matching bias was at 0.01. Including an unobserved site-level covariate to the assignment mechanism did not significantly increase ATT bias for any of the methods (Figure 5.6, panel c), although performance of the two-stage matching method still worsened with the more flexible propensity score models. Additionally, pooled matching performed surprisingly well, particularly with a RIS propensity score model. The same relative patterns existed under an assignment mechanism that also depended on an unobserved unit-level covariate (Figure 5.6, panel d), but, as expected, all methods experienced an increase in bias.

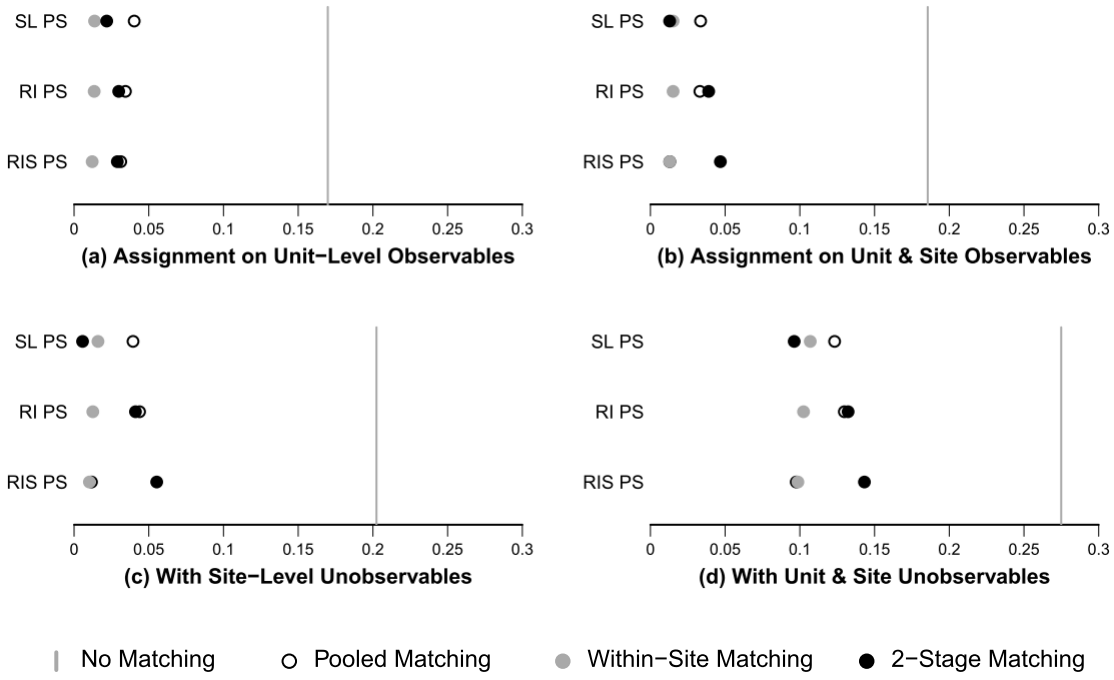

Figure 5.6. Average grand-mean ATT bias across simulation replications, by assignment mechanism and design phase conditions

Bias in site-level ATT variance estimation was also significantly reduced under each matching conditions (see Figure 5.7), relative to the naïve estimation approach. The true ATT variance was approximately 0.17 to 0.19, depending on the assignment condition. Under the random assignment condition, all matching methods estimated ATT variance about 0.01 of a point above the true value, on average. Given a Monte Carlo standard deviation of approximately 0.03 under all matching conditions, conditioning through matching produced a statistically less biased estimate of ATT variance, even under random assignment. The differences, however, were substantively negligible. When assignment only depended on observed unit-level covariates (Figure 5.7, panel a), ATT variance bias for within-site matching and two-stage matching was at or below 0.01 regardless of the propensity score model. As a percentage of true ATT variance (100 × bias/true ATT variance), these matching methods reduced bias from about 43% under the

naïve estimation method to less than 6%. Bias reduction under pooled matching was slightly worse, with the percent of bias variance around 15%.

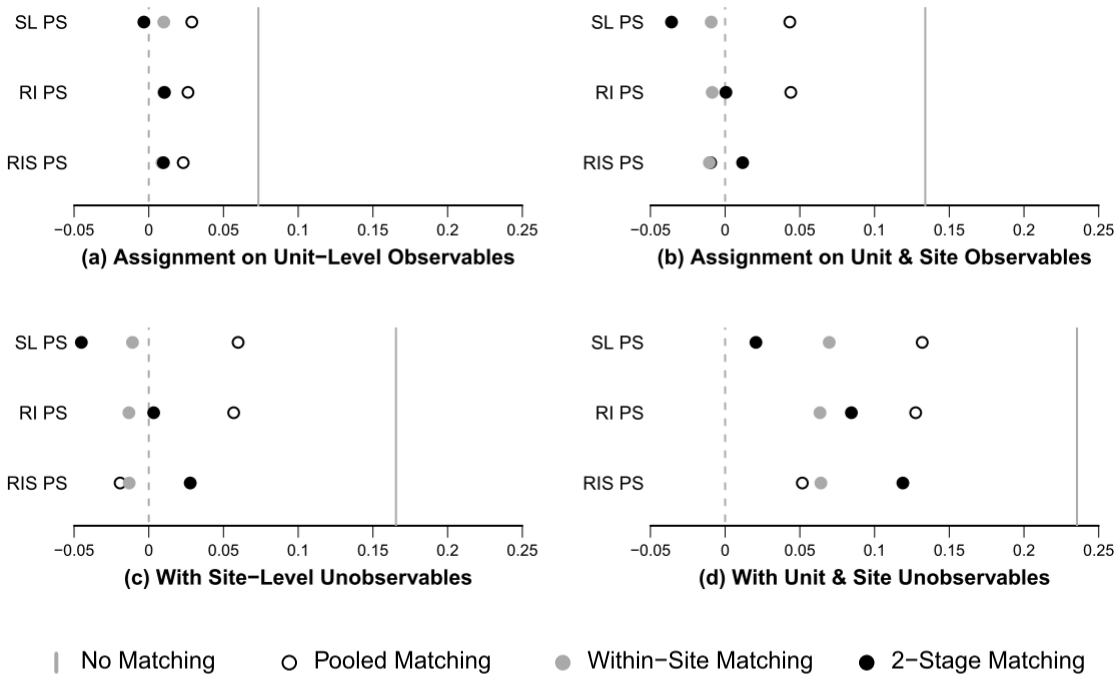

Figure 5.7. Average site-level ATT variance bias across simulation replications, by assignment mechanism and design phase conditions

When assignment also depended on an observed site-level covariate (Figure 5.7, panel b), within-site matching consistently performed well regardless of the propensity score model, while two-stage matching under-estimated variance with a single-level propensity score model (bias = -0.04) but performed similarly to within-site matching when based on a RI or RIS model. Pooled matching only performed well when paired with a RIS model. These differences were magnified when assignment also depended on unobserved covariates (Figure 5.7, bottom row).

In summary, these simulation results show the strength of within-site matching to reduce bias in both the grand-mean ATT and site-level variance estimates, regardless of the propensity

score model. Performance of pooled matching and two-stage matching, on the other hand, depended on the propensity score model employed. While pooled matching performed surprisingly well with a RIS model, grand-mean ATT bias under two-stage matching was minimized with a single-level propensity score model. However, use of a single-level propensity score model with two-stage matching under-estimated ATT variance compared to matching with a RI or RIS model. The trend in two-stage matching under different propensity score models mirrors the simulation results regarding covariate balance, so conditions with less balance reduction translated into less bias reduction in the effect estimates.

One possible explanation for why two-stage matching performed worse under the more flexible propensity score model might have to do with how two-stage matching uses the propensity score model in the between-site matching stage. Both the propensity score model predicted values and the parameter estimates are estimated with error. None of the methods incorporate this estimation uncertainty in the analysis, but pooled matching and within-site matching only rely on the predicted propensity score values. The predicted values are generally measured with more precision, given that parameter estimation error can balance out across the model, so error in the predicted values is less likely to affect matching results. Two-stage matching, however, depends on both the predicted values and the model parameter estimates because the predicted scores are recalculated in the between-site matching stage based on the treatment site's model parameter estimates. Given limited within-site sample size, the more flexible propensity score models will tend to have more parameter error and this error could result in more biased effect estimates.

5.3. How adjustment phase specifications influence treatment effect estimates

In the previous chapter, the empirical illustration suggested that the overall average 8th grade algebra effect was slightly smaller when the adjustment phase was omitted but did not differ substantively when based on a single imputation or five multiple imputations. The estimate of between-site effect variance was also stable across changes in the number of imputations. In this section, I present results from the simulation study that formally tested differences in the number of adjustments. As described in Chapter 3, four different conditions were tested in this simulation study:

- No adjustment in between-site matched control units
- One imputed adjustment for between-site effect differences
- Five imputed adjustments for between-site effect differences
- Ten imputed adjustments for between-site effect differences

Under each of the five assignment mechanism conditions, the grand-mean ATT and site-level ATT variance were estimated under the different adjustment phase conditions. I used the two-stage matching method with a RIS propensity score model in the design phase and an unconditional two-level HLM for effect estimation in the analysis phase. I examined the same simulation summary statistics as in the design phase simulation study. The results are presented in Table 5.1.

Table 5.1. Adjustment phase simulation study results summarized across replications, by assignment mechanism and adjustment phase condition.

Assign. Mech.	Adjust Method	Grand-Mean ATT					Site-Level Effect Variance		
		Bias	(MCSD)	Std. Err.	RMSE	Cover.	Bias	(MCSD)	RMSE
RA	None	-0.03	(0.06)	0.06	0.07	0.94	0.02	(0.03)	0.03
	1 Imp	-0.03	(0.06)	0.06	0.07	0.94	0.02	(0.03)	0.03
	5 Imp	-0.03	(0.06)	0.06	0.07	0.94	0.02	(0.03)	0.03
	10 Imp	-0.03	(0.06)	0.06	0.07	0.94	0.02	(0.03)	0.03
L1OB	None	0.00	(0.06)	0.06	0.06	0.98	0.03	(0.02)	0.04
	1 Imp	0.00	(0.06)	0.06	0.06	0.97	0.01	(0.02)	0.03
	5 Imp	0.00	(0.06)	0.06	0.06	0.97	0.01	(0.02)	0.03
	10 Imp	0.00	(0.06)	0.06	0.06	0.97	0.01	(0.02)	0.03
L2OB	None	0.01	(0.06)	0.07	0.06	0.98	0.07	(0.03)	0.07
	1 Imp	0.02	(0.06)	0.07	0.06	0.96	0.02	(0.03)	0.04
	5 Imp	0.02	(0.06)	0.07	0.06	0.96	0.02	(0.03)	0.04
	10 Imp	0.02	(0.06)	0.07	0.06	0.96	0.02	(0.03)	0.04
L2UN	None	0.06	(0.07)	0.08	0.09	0.93	0.13	(0.04)	0.14
	1 Imp	0.03	(0.07)	0.07	0.07	0.95	0.04	(0.03)	0.05
	5 Imp	0.03	(0.07)	0.07	0.07	0.94	0.04	(0.03)	0.05
	10 Imp	0.03	(0.07)	0.07	0.07	0.95	0.04	(0.03)	0.05
L1UN	None	0.15	(0.07)	0.09	0.16	0.65	0.22	(0.04)	0.22
	1 Imp	0.12	(0.07)	0.08	0.13	0.71	0.13	(0.03)	0.13
	5 Imp	0.12	(0.07)	0.08	0.13	0.71	0.13	(0.03)	0.13
	10 Imp	0.12	(0.07)	0.08	0.13	0.71	0.13	(0.03)	0.13

Notes: RA = random assignment; L1OB = unit-level observed covariates; L2OB = unit- & site-level observed covariates; L2UN = site-level unobserved covariate; L1UN = unit- & site-level unobserved covariates. MCSD = Monte Carlo standard deviation; RMSE = root mean squared error.

The simulation study results support the empirical illustration finding regarding effect robustness across the number of imputations used in the adjustment phase (see Table 5.1). When treatment assignment was random or only depended on observed covariates, the decision to adjust or not adjust the between-site matched control unit outcomes did not influence relative bias in the grand-mean ATT. Bias in the site-level effect variance estimate, when assignment

depended on a site-level observed covariate, was reduced, however, when at least one adjustment was made. As a percentage of true between-site ATT variance, variance bias without the adjustment was about 40%, but was only 11% after an adjustment based on one imputation for school-effects. Increasing the number of imputations did not change the amount of remaining bias. A similar pattern existed when assignment depended on unobserved covariates. This suggests that the use of between-site matching can introduce bias into the matched groups for some sites, but at least some of that bias can be extracted in the adjustment phase. Based on both the simulation results and the empirical illustration, the added complexity associated with multiply imputing site effects for the adjustments is not justified given insignificant changes in the effect estimates across different number of imputations.

5.4. How analysis phase specifications influence treatment effect estimates

For the design phase and adjustment phase simulation studies, the effect estimates were based on an unconditional two-level model. In practice, a variety of different models could be used in the analysis phase for treatment effect estimation. Furthermore, combining matching with regression-model covariate adjustment is recommended in the literature (Ho et al., 2007; Schafer & Kang, 2008) as a way to decrease bias and improve precision. So while the two-stage matching method may not result in perfect covariate balance, or produce unbiased effect estimates under an unconditional outcome model, it may prove effective with a conditional outcome model that controls for residual covariate imbalance. In this section, I present results from the simulation study that formally tested whether treatment effect estimation is sensitive to different outcome model specifications. As described in Chapter 3 (see Table 3.3), five different outcome model specifications were tested in this simulation study:

- Single-level model with treatment group indicator (SL UN);
- Single-level model with treatment group indicator and controls for observed unit- and site-level covariates (SL CN);
- Two-level RIS model with treatment group indicator at level-1 (ML UN);
- Two-level RIS model with treatment group indicator and controls for observed unit-level covariates (ML CN1);
- Two-level RIS model with treatment group indicator and controls for observed unit- and site-level covariates (ML CN2).

Under each of the five assignment mechanism conditions, the grand-mean ATT and site-level ATT variance were estimated under the different model specification conditions. For each model specification, the grand-mean ATT and ATT variance were estimated using the unmatched data and the data preprocessed with the two-stage matching approach. Note, however, that the single-level models treat the average treatment effect as fixed and therefore do not estimate between-site variance. Additionally, the between-site variance estimate from the two-level RIS model with level-2 controls (ML CN2) is not the ATT variance, but residual effect variance after accounting for effect heterogeneity in the level-2 confounder. So ATT variance estimates are only compared across two model specifications: ML UN and ML CN1. I examined the same simulation study summary statistics as in the design phase simulation study. The full simulation results are presented in Table A.9 in the appendix.

For estimation of the grand-mean ATT, bias was significantly reduced for both the unmatched and matched data when a conditional multilevel model was employed (see Figure 5.8). Not surprisingly, estimation based on the unmatched data was more sensitive to model specification than estimation based on the matched data. For example, average grand-mean ATT

bias when assignment depended on observed unit- and site-level covariates (Figure 5.8, panel b) ranged from -0.05 to 0.57 for model estimates based on the unmatched data, but only ranged from -0.01 to 0.17 for model estimates based on the matched data. For both the unmatched and matched condition, however, bias was minimized—and similar—when a conditional multilevel model was employed. This was true when assignment only depended on observed covariates and when assignment depended on observed and unobserved covariates. Combining multilevel regression-based covariate adjustment with the two-stage matching approach facilitated unbiased ATT effect estimation (bias with 0.01 of the true grand-mean ATT) under all the assignment mechanisms except for assignment dependent on an unobserved unit-level covariate. This marked a significant reduction in bias relative to estimation based on two-stage matching and an unconditional multilevel model (e.g., from 0.05 to -0.01 when assignment depended on observed unit- and site-level covariates).

Use of a conditional instead of an unconditional multilevel model significantly improved estimation of between-site ATT variance for the unmatched data, but not for the matched data (see Figure 5.9). For example, when assignment depended on observed unit- and site-level covariates, bias in the ATT variance estimate using unmatched data was about 79% of the true variance with an unconditional model but only 19% of the true variance with a conditional model. By comparison, bias in the ATT variance estimate using the matched data was 7% of true variance with an unconditional model and 9% of true variance with a conditional model. A similar pattern existed when assignment also depended on an unobserved site-level covariate. These results suggest that even when one uses a conditional multilevel model to estimate between-site ATT variance, first preprocessing the data with the two-stage matching approach

will produce a less biased estimate and the estimate will be less sensitive to outcome model specification.

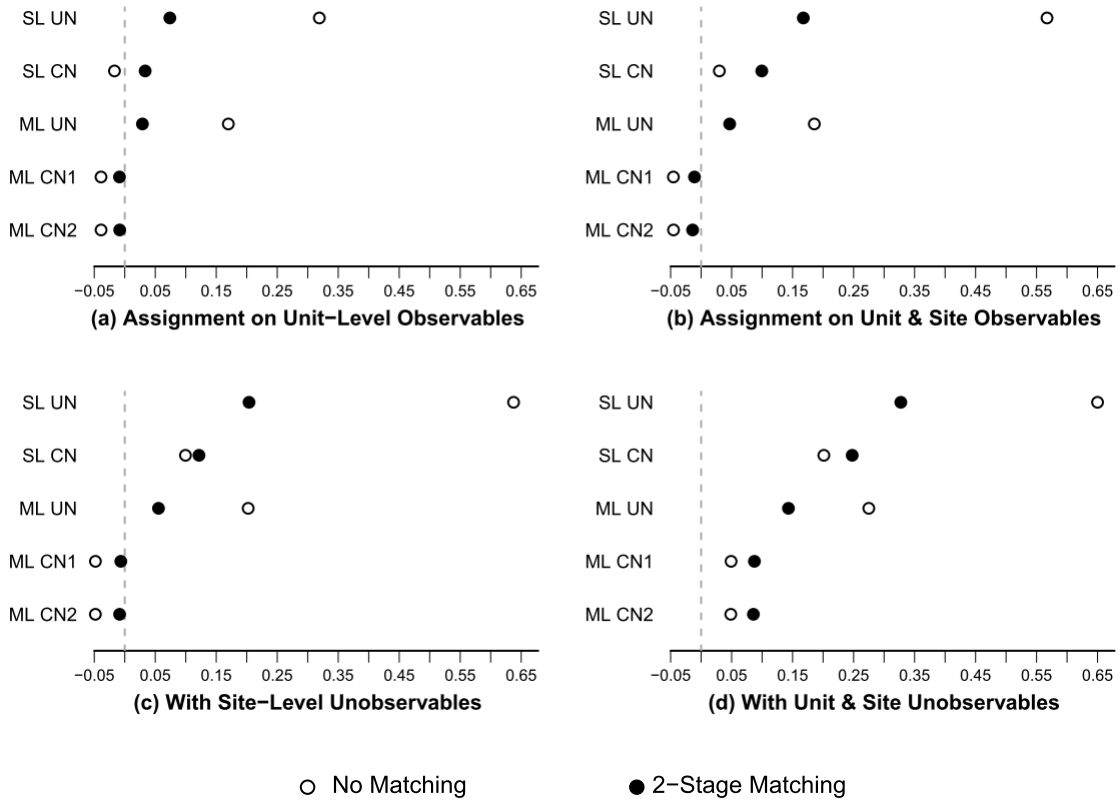

Figure 5.8. Average grand-mean ATT bias across simulation replications, by assignment mechanism and analysis phase conditions

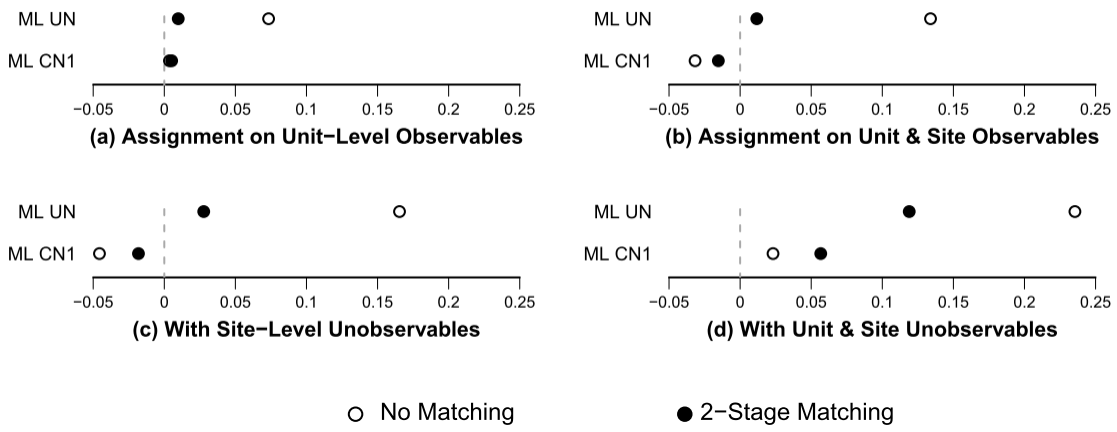

Figure 5.9. Average between-site ATT variance bias across simulation replications, by assignment mechanism and analysis phase conditions

5.5. Summary of key findings from the simulation studies

In this chapter I presented results from simulation studies designed to test the performance of alternative specifications to the proposed two-stage matching strategy and gauged its performance relative to other effect estimation approaches. In general, the findings support results from the empirical illustration regarding similarity of results across estimation methods, but also suggest certain conditions under which the two-stage matching approach performs well. In the design phase, all matching methods improved covariate balance. While the two-stage matching method did not outperform within-site matching and only performed marginally better than pooled matching in terms of covariate balance, more treatment units were retained in the analysis with the two-stage method. Similarly, all matching method options tested in the design phase significantly reduced bias in treatment effect estimates. Pooled matching performed surprisingly well when paired with a RIS propensity score model, while two-stage matching performed best when paired with a single-level propensity score model. In the adjustment phase, making an adjustment for site-effect differences in the between-site matched

control units reduced bias in effect estimates, but no discernible improvement came from multiply imputing the adjustment. In the analysis phase, combining multilevel regression-based covariate adjustment with the two-stage matching approach facilitated unbiased ATT effect estimation.

The simulation studies were limited in scope given a desire to focus on key aspects of the proposed method and time limitations. Future work should look to extend the simulations to examine performance of the two-stage matching strategy under different conditions. In particular, the simulations were based on a fixed sample size of 50 sites and an average of 200 units per site. This sample size was selected to represent data from a modest real-world study. I hypothesize that the two-stage matching method will be more effective as sample size (at the site- and/or unit-levels) increases. Additionally, all propensity score matching in the simulation studies was executed using 1-to-1 within-caliper matching, with a caliper of 0.25 of a standard deviation. I chose this type of matching because it represents one of the most common forms of propensity score matching and is the method used by Stuart and Rubin (2008). Stuart and Rubin showed how increasing the caliper range prioritizes within-site matching over between-site matching, while lowering the caliper range does the opposite. Future research could examine sensitivity of the two-stage matching strategy to changes in the caliper range, as well as how exploration of the data can inform where to set the caliper range. Other aspects of the two-stage matching approach that warrant future investigation include alternative ways to specify site clusters in the design phase and alternative ways to estimate site effects in the adjustment phase. Additionally, the simulation studies should be expanded to test the two-stage matching method's relative performs for estimating treatment-by-covariate interaction effects, since the empirical illustration results indicated these estimates are sensitive to the matching method employed.

Chapter 6

Summary and Discussion

The use of multilevel models is now commonplace in educational research to study educational phenomenon within the education system's hierarchical, or multilevel, structure. Similarly, causal inference under the potential outcomes framework is becoming standard in educational research aimed at connecting policies, programs, and practices to outcomes. Understanding the nexus of these two methodological advances is still in its infancy, however. Literature on causal inference often ignores, or avoids, discussion of a multilevel context, and literature on multilevel modeling often ignores, or avoids, discussion of causal inference outside of a randomized control trial. Despite the increased emphasis on randomized experimental research designs in education over the past decade (Shadish & Cook, 2009), rapidly expanding and improving educational data infrastructures at the state and school district levels (Means, Padilla, & Gallagher, 2010) also provide increased opportunities to examine a multitude of educational questions through non-experimental designs.

Researchers seeking to draw causal inferences from these non-experimental multilevel data must understand how the multilevel structure can complicate causal effect estimation and how multilevel models and matching methods can aid the estimation effort. To date, the application of multilevel modeling and propensity score matching under a potential outcomes framework is relegated to a relatively small set of articles (Arpino & Mealli, 2011; Hong & Raudenbush, 2006; Kelcey, 2011a, 2011b; Kim & Seltzer, 2007; Rosenbaum, 1986; Steiner, 2011; Su & Cortina, 2009; Thoemmes, 2009; Thoemmes & West, 2011). This study adds to the

research methodology literature in a number of ways. In this concluding chapter I discuss these methodological contributions, along with potential areas for future research.

6.1. Complications and considerations for causal inference in a multilevel setting

Through both the literature review and the analysis, I highlighted some of the key complications and considerations one must face when trying to draw causal inferences in a multilevel setting. For example, for the strongly ignorable treatment assignment assumption to hold in a multisite setting, the research must consider both unit- and site-level confounders as well as potential cross-level interactions. Additionally, conceptualizing the analysis as a series of within-site mini-studies and using multilevel modeling to pool the estimates across sites (Seltzer, 2004) can allow one to relax SUTVA (Gitelman, 2005). Relying solely on a multilevel regression model to properly account for pretreatment covariate group differences for unbiased effect estimation, however, places a lot of faith in the model's parameterization and extrapolation. Therefore, preprocessing the data through matching is often desirable to limit covariate bias prior to effect estimation.

The simulation study findings reinforced a few key notions pertaining to causal effect estimation. First, given finite sample sizes, researchers need to attend to covariate imbalance even if the treatment is randomly assigned. This is particularly true in a multisite setting where one wants to balance treatment and control groups within each site. Stratifying on key covariates prior to randomization or using ANCOVA after randomization may help overcome this potential problem. Second, preprocessing the data via matching significantly reduces covariate imbalance and combining matching with a covariate adjustment regression-based approach, i.e., dual modeling, is better than relying on one approach in isolation (Ho et al., 2007; Schafer & Kang,

2008; Su & Cortina, 2009). Ultimately, however, the utility of any covariate adjustment method depends on whether the correct confounders are included in the analysis. The simulation results showed that under selection on the observables, average treatment effect estimates were close to the true average effect, but when selection also depended on an unobserved factor—particularly a unit-level factor—all estimation approaches resulted in biased estimates. This finding is consist with other studies that compared performance of different effect estimation methods (Cook et al., 2008; Cook, Steiner, & Pohl, 2009; Schafer & Kang, 2008; Shadish et al., 2008), and highlights the importance of investigating the assignment process to identify the important confounding factors (Rickles, 2011).

6.2. Two-stage matching: an alternative to within-site matching and pooled matching

I proposed, demonstrated, and tested a matching method that extends Stuart and Rubin's (2008) multiple control group strategy to a multilevel setting. The two-stage matching method provides researchers with an alternative to strict within-site matching or pooled matching. While within-site matching is ideal when trying to approximate a multisite design (Rosenbaum, 1986), the method can be limited by sample size restrictions and within-site covariate overlap (Kelcey, 2011a, 2011b; Thoemmes & West, 2011). Conversely, pooled matching lacks the appealing quality of matching on both observed and unobserved site-level covariates, where matches may, in the language of Shadish and Cook (2009), be focal but not local. By first prioritizing within-site matching in stage one and supplementing within-site matches with between-site matching in stage two, the two-stage matching method balances the strengths and weaknesses of within-site and pooled matching.

The two-stage matching method has conceptual appeal because it emphasizes study design, tries to approximate a multisite randomized design, and acknowledges the desire for focal local controls. The method can also retain more treatment units in the matched analysis, which might have implications for generalizability. Generalizability is a particular concern given effect heterogeneity because treatment effects based on an analysis that excludes certain units (e.g., units with high propensity scores) and/or sites (e.g., sites with very selective treatment assignment) might not reflect the true average treatment effect in the population, nor provide a satisfactory way to investigate effect heterogeneity. However, effect estimation under the two-stage matching method does not outperform more traditional matching-based or regression-based methods under most conditions. Given the relatively small differences across estimation options found in the empirical illustration and simulation studies, researchers may not find the added complexity of the two-stage matching method worthwhile if the only payoff is conceptual peace of mind.

The simulation results do suggest that the two-stage matching approach can perform well with some simplifying alternative specifications. In the design phase, two-stage matching performance improved with a single-level propensity score model instead of a more computationally demanding RIS multilevel model. In the adjustment phase, two-stage matching performance improved when at least one adjustment was made for site effect differences introduced by the between-site matching. Effect estimation based on multiply imputing site effects was not substantially different from estimation based on a single imputation, however. In the analysis phase, performance of the two-stage matching method, as with the more traditional methods, improved when paired with a conditional multilevel outcome model instead of an

unconditional model. Furthermore, the two-stage matching method may prove to be a useful option for investigating sources of effect heterogeneity both among units and between sites.

6.3. Concluding implications for multisite observation studies

In addition to extending the work of Stuart and Rubin, this study complements the small set of studies that examined how to apply propensity score matching in multisite settings (Arpino & Mealli, 2011; Hong & Raudenbush, 2006; Kelcey, 2011b; Su & Cortina, 2009; Thoemmes, 2009; Thoemmes & West, 2011). Paired with past research, the study's findings provide some guidance for researchers looking to estimate treatment effects from a multisite observational study. Most importantly, researchers should make efforts to understand what factors influence both treatment assignment and outcomes of interest, giving particular attention to the importance of site-level factors relative to unit-level factors. Determining where site-level factors fall in reference to Cochran's (1965) three classes of disturbing variables can inform the type of matching to use. When site-level factors play an important role in treatment assignment, within-site matching is preferred. The two-stage matching method provides an alternative option when within-site matching is limited, and can perform well under more simplified specifications (e.g., single-level propensity score model and single imputation of site-effects). The two-stage matching method may be particularly effective and efficient if site-level factors fall within Cochran's second class of disturbing variables: variables we would like to match but their effects produce little bias. If the researcher prefers, or is limited to, pooled matching, it should be conducted with a multilevel propensity score model that accounts for heterogeneity in treatment assignment across sites. When site-level factors play an insignificant role in treatment assignment and/or outcomes (Cochran's third class of disturbing variables), pooled matching

should be preferred since site-level factors can be ignored. Regardless of the propensity score matching method, average treatment effects and effect heterogeneity should be estimated based on the matched data and multilevel outcome models that adjust for remaining covariate imbalance. Sensitivity analysis is also recommended to examine how sensitive findings are to potentially unobserved confounders.

6.4. Directions for future research

In demonstrating and testing the proposed two-stage matching method for this study, I was not able to address some important methodological components that warrant investigation. I touched on some of the desired areas for future research at the end of Chapter 5, including performance of two-stage matching under different sample size conditions, matching algorithms, and site cluster definitions. In the design phase, an unexpected finding from the simulation study was how two-stage matching performed better under a single-level propensity score model instead of the RIS model. Future research should seek to understand how error in the estimated propensity score model parameters influence performance of the two-stage matching method, whether this error explains the method's worse performance with a RIS model, and whether efforts to account for this error—possibly through Bayesian modeling (Kaplan & Chen, 2011)—can improve performance. Additionally, future work could examine whether, given a RIS propensity score model, the proposed recalibration of the predicted propensity score for the between-site matching is efficient or necessary.

In the adjustment phase, alternative methods for site effect estimation and multiple imputation of effects should be examined. First, I based site effect estimation on all control units that had a predicted propensity score within the matched control group range. This differs from

the Stuart and Rubin (2008) approach, where they based site effect estimation on a subsample of matched control units. Understanding how site effect estimates change based on different control unit samples is a worthwhile undertaking. Second, the site effect model used to demonstrate and test the two-stage matching method could either be inefficiently complicated or ineffectively simple. I chose to use a non-uniform, or heterogeneous, site effect model (Raudenbush & Willms, 1995; Reardon & Raudenbush, 2009) where site effects are a function of both mean differences and a linear relationship between a key covariate and the outcome. The method may perform as well based on a uniform, or homogeneous, site effects model. On the other hand, it may be necessary to control for and/or allow site effects to differ across more than one covariate. Future work could examine whether two-stage matching performance is sensitive to changes in the site effects model. Additionally, I opted for one of many options for multiply imputing site effects. Alternative imputation methods, particularly Markov chain Monte Carlo methods, could be tested.

The empirical illustration also raised important considerations researchers may have to address if seeking to implement the two-stage matching method in practice. For the illustration, the data set was treated as if it did not contain missing data. In practice, researchers will have to address how to handle missing data before executing two-stage matching. If missing data are multiply imputed and school effects in the adjustment phase are multiply imputed, then data analysis will have to incorporate uncertainty in both of the missing data imputation and site effect estimation. Similarly, some students were missing the outcome value because of attrition (e.g., dropping out of school prior to 10th grade). If this attrition is associated with treatment assignment, then the effect estimates might be biased. Future research must examine ways to address these missing data shortcomings within the matching analytic framework.

Demonstration of the proposed method also ignored the fact that the matching process can result in some control units being matched to multiple treatment units. Future research should investigate how matching with replacement in the between-site matching stage impacts effect estimates and standard errors, as well as what options—such as weighting—can appropriately account for repeated control units. Additionally, the empirical illustration and simulation studies demonstrated and tested the two-stage matching method as it pertains to a continuous outcome measure. The two-stage matching method should be extended to binary outcomes, such as whether a student graduated from high school or not. Extending the method to binary, or ordinal, outcomes is straightforward in the design and analysis phases, but it is not clear how one should estimate site effects in the adjustment phase if the outcome is not continuous.

Lastly, the empirical illustration raises a methodological concern that is not unique to the two-stage matching method, and warrants consideration in future research. For the empirical illustration I focused on estimating the effect of a treatment assigned when students were in middle school on an outcome measured when students were in high school. The proposed method focuses on estimating this effect given that students are nested within middle schools at the time of treatment assignment. Over the time period between treatment assignment and the outcome measurement, however, students became cross-classified in middle schools and high schools. To my knowledge no research on propensity score matching, or the broader causal inference literature, addresses how to handle this type of cross-classified data. On one hand, not accounting for the fact that students are nested within high schools as well as middle schools could result in underestimated standard errors. On the other hand, nesting within high schools occurs after treatment assignment and can therefore be thought of as a post-treatment factor.

Conditioning on any post-treatment factor could result in biased effect estimates, particularly if treatment assignment and high school selection are related (Frangakis & Rubin, 2002). While researchers continue to wrestle with methods for drawing causal inferences in hierarchical settings, attention will also have to turn toward even more complicated settings such as the existence of cross-classified data.

Appendix

Detailed Tables with Simulation Study Results

The appendix includes detailed summary statistics for the simulation study results discussed in Chapter 5. The tables include a variety of abbreviations/acronyms, which are defined below.

Summary Statistics:

ASB = absolute standardized bias

VR = variance ratio

MCSD = Monte Carlo standard deviation

Site Var = between-site variance in target statistic

Site Max = maximum site-level value in target statistic

ATT = average treatment effect for the treated

Std. Err. = mean standard error of the ATT estimate across Monte Carlo replications

RMSE = root mean squared error

Cover. = approximate 95% coverage rate

Assignment Mechanism:

RA = random assignment

L1OB = selection on unit-level observables

L2OB = selection on unit- and site-level observables

L2UN = selection on unit-level observables and site-level observables and unobservables

L1UN = selection on unit- and site-level observables and unobservables

Covariates (CV):

PS = predicted propensity score

X = observed unit-level covariate

Z = observed unit-level covariate

U = unobserved unit-level covariate

S = observed site-level covariate

V = unobserved site-level covariate

Propensity Score Model Condition:

SL PS = single-level propensity score model

RI PS = random intercept two-level propensity score model

RIS PS = random-intercept-and-slope two-level propensity score model

Matching Method Condition:

PM = pooled matching

WM = within-site matching

2SM = two-stage matching

Analysis Model Condition:

SL UN = single-level unconditional model

SL CN = single-level conditional model

ML UN = multilevel unconditional model

ML CN1 = multilevel model conditional on observed unit-level covariates

ML CN2 = multilevel model conditional on observed unit- and site-level covariates

Table A.1. Design phase simulation study results summarized across replications: Level-1 covariate balance under random assignment (RA)

	PS	Match	Within-Site ASB				Within-Site VR			
CV	Model	Method	Mean	(MCSD)	Site Var	Site Max	Mean	(MCSD)	Site Var	Site Max
PS	SL PS	None	0.12	(0.01)	0.01	0.39	1.01	(0.03)	0.05	1.58
		PM	0.13	(0.01)	0.01	0.42	1.04	(0.03)	0.06	1.74
		WM	0.04	(0.00)	0.00	0.10	1.01	(0.01)	0.00	1.11
		2SM	0.04	(0.00)	0.00	0.11	1.02	(0.01)	0.00	1.14
	RI PS	None	0.12	(0.01)	0.01	0.39	1.01	(0.03)	0.05	1.58
		PM	0.13	(0.01)	0.01	0.43	1.04	(0.02)	0.06	1.71
		WM	0.04	(0.00)	0.00	0.10	1.01	(0.01)	0.00	1.11
		2SM	0.05	(0.04)	0.01	0.20	1.00	(0.08)	0.01	1.17
	RIS PS	None	0.13	(0.02)	0.01	0.42	1.01	(0.03)	0.05	1.59
		PM	0.12	(0.01)	0.01	0.37	1.04	(0.03)	0.06	1.71
		WM	0.05	(0.00)	0.00	0.10	1.02	(0.01)	0.00	1.12
		2SM	0.06	(0.02)	0.00	0.26	1.01	(0.04)	0.02	1.29
X	None	None	0.12	(0.01)	0.01	0.38	1.01	(0.03)	0.05	1.58
	SL PS	PM	0.13	(0.01)	0.01	0.43	1.03	(0.03)	0.06	1.69
		WM	0.10	(0.03)	0.01	0.30	1.02	(0.03)	0.04	1.55
		2SM	0.09	(0.03)	0.01	0.30	1.03	(0.03)	0.04	1.58
	RI PS	PM	0.13	(0.01)	0.01	0.42	1.03	(0.03)	0.06	1.72
		WM	0.10	(0.03)	0.01	0.31	1.02	(0.03)	0.04	1.57
		2SM	0.10	(0.03)	0.01	0.31	1.03	(0.03)	0.04	1.56
	RIS PS	PM	0.12	(0.02)	0.01	0.38	1.03	(0.03)	0.06	1.72
		WM	0.09	(0.02)	0.01	0.32	1.03	(0.03)	0.04	1.63
		2SM	0.09	(0.02)	0.01	0.32	1.04	(0.03)	0.04	1.63
Z	None	None	0.12	(0.01)	0.01	0.38	1.01	(0.03)	0.04	1.57
	SL PS	PM	0.13	(0.01)	0.01	0.42	1.04	(0.03)	0.06	1.72
		WM	0.09	(0.03)	0.01	0.29	1.02	(0.03)	0.04	1.51
		2SM	0.09	(0.03)	0.01	0.28	1.03	(0.03)	0.04	1.53
	RI PS	PM	0.13	(0.01)	0.01	0.42	1.03	(0.03)	0.06	1.72
		WM	0.09	(0.03)	0.01	0.28	1.02	(0.03)	0.04	1.50
		2SM	0.09	(0.03)	0.01	0.28	1.03	(0.03)	0.04	1.52
	RIS PS	PM	0.12	(0.01)	0.01	0.38	1.03	(0.03)	0.06	1.68
		WM	0.08	(0.02)	0.00	0.30	1.02	(0.03)	0.04	1.58
		2SM	0.08	(0.02)	0.00	0.30	1.03	(0.03)	0.04	1.56
U	None	None	0.12	(0.01)	0.01	0.38	1.02	(0.03)	0.05	1.60
	SL PS	PM	0.13	(0.01)	0.01	0.42	1.03	(0.04)	0.06	1.70
		WM	0.13	(0.01)	0.01	0.42	1.03	(0.04)	0.06	1.72
		2SM	0.13	(0.01)	0.01	0.41	1.03	(0.04)	0.06	1.71
	RI PS	PM	0.13	(0.01)	0.01	0.41	1.03	(0.04)	0.06	1.73
		WM	0.13	(0.02)	0.01	0.43	1.03	(0.04)	0.06	1.72
		2SM	0.13	(0.01)	0.01	0.42	1.02	(0.04)	0.06	1.71
	RIS PS	PM	0.13	(0.01)	0.01	0.42	1.03	(0.04)	0.06	1.73
		WM	0.13	(0.01)	0.01	0.42	1.03	(0.04)	0.06	1.72
		2SM	0.13	(0.01)	0.01	0.42	1.03	(0.04)	0.06	1.71

Table A.2. Design phase simulation study results summarized across replications: Level-1 covariate balance under treatment assignment with level-1 observed covariates (L1OB)

	PS	Match	Within-Site ASB				Within-Site VR			
CV	Model	Method	Mean	(MCSD)	Site Var	Site Max	Mean	(MCSD)	Site Var	Site Max
PS	SL PS	None	0.44	(0.02)	0.02	0.80	1.02	(0.03)	0.05	1.62
		PM	0.15	(0.01)	0.01	0.46	1.14	(0.03)	0.07	1.89
		WM	0.07	(0.01)	0.00	0.16	1.06	(0.01)	0.00	1.18
		2SM	0.06	(0.00)	0.00	0.14	1.04	(0.01)	0.01	1.26
	RI PS	None	0.44	(0.02)	0.02	0.80	1.02	(0.03)	0.05	1.62
		PM	0.14	(0.01)	0.01	0.44	1.15	(0.02)	0.07	1.93
		WM	0.07	(0.01)	0.00	0.16	1.06	(0.01)	0.00	1.18
		2SM	0.07	(0.01)	0.00	0.18	1.06	(0.01)	0.00	1.24
	RIS PS	None	0.45	(0.02)	0.02	0.81	1.02	(0.03)	0.05	1.62
		PM	0.13	(0.01)	0.01	0.42	1.15	(0.03)	0.07	1.90
		WM	0.07	(0.01)	0.00	0.16	1.06	(0.01)	0.00	1.18
		2SM	0.08	(0.01)	0.00	0.19	1.07	(0.01)	0.01	1.29
X	None	None	0.34	(0.02)	0.02	0.70	1.02	(0.03)	0.05	1.59
	SL PS	PM	0.14	(0.01)	0.01	0.44	1.09	(0.03)	0.07	1.83
		WM	0.10	(0.01)	0.01	0.31	1.06	(0.03)	0.05	1.66
		2SM	0.09	(0.01)	0.01	0.30	1.04	(0.03)	0.04	1.64
	RI PS	PM	0.14	(0.01)	0.01	0.44	1.10	(0.03)	0.07	1.80
		WM	0.10	(0.01)	0.01	0.31	1.05	(0.03)	0.05	1.66
		2SM	0.10	(0.01)	0.01	0.30	1.05	(0.03)	0.04	1.61
	RIS PS	PM	0.13	(0.01)	0.01	0.41	1.10	(0.03)	0.07	1.83
		WM	0.09	(0.01)	0.00	0.28	1.05	(0.03)	0.05	1.66
		2SM	0.09	(0.01)	0.00	0.27	1.05	(0.03)	0.04	1.62
Z	None	None	0.27	(0.02)	0.02	0.63	1.02	(0.03)	0.04	1.59
	SL PS	PM	0.14	(0.01)	0.01	0.44	1.07	(0.03)	0.06	1.78
		WM	0.11	(0.01)	0.01	0.36	1.05	(0.03)	0.06	1.71
		2SM	0.11	(0.01)	0.01	0.35	1.03	(0.03)	0.05	1.67
	RI PS	PM	0.13	(0.01)	0.01	0.43	1.07	(0.03)	0.06	1.75
		WM	0.11	(0.01)	0.01	0.36	1.05	(0.03)	0.06	1.72
		2SM	0.11	(0.01)	0.01	0.35	1.04	(0.03)	0.05	1.66
	RIS PS	PM	0.13	(0.01)	0.01	0.41	1.07	(0.03)	0.06	1.78
		WM	0.11	(0.01)	0.01	0.33	1.05	(0.03)	0.06	1.73
		2SM	0.10	(0.01)	0.01	0.32	1.04	(0.03)	0.05	1.68
U	None	None	0.12	(0.01)	0.01	0.38	1.02	(0.03)	0.05	1.60
	SL PS	PM	0.13	(0.02)	0.01	0.43	1.03	(0.04)	0.06	1.74
		WM	0.14	(0.01)	0.01	0.44	1.04	(0.04)	0.07	1.76
		2SM	0.13	(0.01)	0.01	0.44	1.02	(0.04)	0.06	1.72
	RI PS	PM	0.13	(0.01)	0.01	0.43	1.03	(0.04)	0.06	1.72
		WM	0.13	(0.01)	0.01	0.45	1.04	(0.04)	0.07	1.77
		2SM	0.13	(0.01)	0.01	0.44	1.03	(0.04)	0.06	1.71
	RIS PS	PM	0.13	(0.02)	0.01	0.44	1.03	(0.03)	0.06	1.74
		WM	0.13	(0.01)	0.01	0.44	1.03	(0.04)	0.06	1.75
		2SM	0.13	(0.02)	0.01	0.44	1.02	(0.03)	0.06	1.71

Table A.3. Design phase simulation study results summarized across replications: Level-1 covariate balance under treatment assignment with level-2 observed covariate (L2OB)

	PS	Match	Within-Site ASB				Within-Site VR			
CV	Model	Method	Mean	(MCSD)	Site Var	Site Max	Mean	(MCSD)	Site Var	Site Max
PS	SL PS	None	0.44	(0.02)	0.02	0.80	1.02	(0.03)	0.05	1.62
		PM	0.15	(0.01)	0.01	0.46	1.14	(0.03)	0.07	1.89
		WM	0.07	(0.01)	0.00	0.16	1.06	(0.01)	0.00	1.18
		2SM	0.06	(0.00)	0.00	0.14	1.04	(0.01)	0.01	1.26
	RI PS	None	0.44	(0.02)	0.02	0.80	1.02	(0.03)	0.05	1.62
		PM	0.14	(0.01)	0.01	0.44	1.15	(0.02)	0.07	1.93
		WM	0.07	(0.01)	0.00	0.16	1.06	(0.01)	0.00	1.18
		2SM	0.07	(0.01)	0.00	0.18	1.06	(0.01)	0.00	1.24
	RIS PS	None	0.45	(0.02)	0.02	0.81	1.02	(0.03)	0.05	1.62
		PM	0.13	(0.01)	0.01	0.42	1.15	(0.03)	0.07	1.90
		WM	0.07	(0.01)	0.00	0.16	1.06	(0.01)	0.00	1.18
		2SM	0.08	(0.01)	0.00	0.19	1.07	(0.01)	0.01	1.29
X	None	None	0.34	(0.02)	0.02	0.70	1.02	(0.03)	0.05	1.59
	SL PS	PM	0.14	(0.01)	0.01	0.44	1.09	(0.03)	0.07	1.83
		WM	0.10	(0.01)	0.01	0.31	1.06	(0.03)	0.05	1.66
		2SM	0.09	(0.01)	0.01	0.30	1.04	(0.03)	0.04	1.64
	RI PS	PM	0.14	(0.01)	0.01	0.44	1.10	(0.03)	0.07	1.80
		WM	0.10	(0.01)	0.01	0.31	1.05	(0.03)	0.05	1.66
		2SM	0.10	(0.01)	0.01	0.30	1.05	(0.03)	0.04	1.61
	RIS PS	PM	0.13	(0.01)	0.01	0.41	1.10	(0.03)	0.07	1.83
		WM	0.09	(0.01)	0.00	0.28	1.05	(0.03)	0.05	1.66
		2SM	0.09	(0.01)	0.00	0.27	1.05	(0.03)	0.04	1.62
Z	None	None	0.27	(0.02)	0.02	0.63	1.02	(0.03)	0.04	1.59
	SL PS	PM	0.14	(0.01)	0.01	0.44	1.07	(0.03)	0.06	1.78
		WM	0.11	(0.01)	0.01	0.36	1.05	(0.03)	0.06	1.71
		2SM	0.11	(0.01)	0.01	0.35	1.03	(0.03)	0.05	1.67
	RI PS	PM	0.13	(0.01)	0.01	0.43	1.07	(0.03)	0.06	1.75
		WM	0.11	(0.01)	0.01	0.36	1.05	(0.03)	0.06	1.72
		2SM	0.11	(0.01)	0.01	0.35	1.04	(0.03)	0.05	1.66
	RIS PS	PM	0.13	(0.01)	0.01	0.41	1.07	(0.03)	0.06	1.78
		WM	0.11	(0.01)	0.01	0.33	1.05	(0.03)	0.06	1.73
		2SM	0.10	(0.01)	0.01	0.32	1.04	(0.03)	0.05	1.68
U	None	None	0.12	(0.01)	0.01	0.38	1.02	(0.03)	0.05	1.60
	SL PS	PM	0.13	(0.02)	0.01	0.43	1.03	(0.04)	0.06	1.74
		WM	0.14	(0.01)	0.01	0.44	1.04	(0.04)	0.07	1.76
		2SM	0.13	(0.01)	0.01	0.44	1.02	(0.04)	0.06	1.72
	RI PS	PM	0.13	(0.01)	0.01	0.43	1.03	(0.04)	0.06	1.72
		WM	0.13	(0.01)	0.01	0.45	1.04	(0.04)	0.07	1.77
		2SM	0.13	(0.01)	0.01	0.44	1.03	(0.04)	0.06	1.71
	RIS PS	PM	0.13	(0.02)	0.01	0.44	1.03	(0.03)	0.06	1.74
		WM	0.13	(0.01)	0.01	0.44	1.03	(0.04)	0.06	1.75
		2SM	0.13	(0.02)	0.01	0.44	1.02	(0.03)	0.06	1.71

Table A.4. Design phase simulation study results summarized across replications: Level-1 covariate balance under treatment assignment with level-2 unobserved covariate (L2UN)

	PS	Match	Within-Site ASB				Within-Site VR			
CV	Model	Method	Mean	(MCSD)	Site Var	Site Max	Mean	(MCSD)	Site Var	Site Max
PS	SL PS	None	0.47	(0.04)	0.09	1.21	1.03	(0.03)	0.05	1.65
		PM	0.26	(0.02)	0.03	0.78	1.00	(0.04)	0.07	1.73
		WM	0.09	(0.01)	0.00	0.23	1.06	(0.01)	0.00	1.20
		2SM	0.07	(0.01)	0.01	0.40	0.97	(0.01)	0.02	1.16
	RI PS	None	0.48	(0.04)	0.09	1.21	1.03	(0.03)	0.05	1.65
		PM	0.27	(0.02)	0.04	0.79	1.00	(0.03)	0.07	1.72
		WM	0.09	(0.01)	0.00	0.23	1.06	(0.01)	0.00	1.20
		2SM	0.15	(0.03)	0.07	1.24	1.01	(0.03)	0.02	1.39
	RIS PS	None	0.48	(0.04)	0.09	1.22	1.03	(0.03)	0.05	1.67
		PM	0.14	(0.01)	0.01	0.43	1.08	(0.03)	0.08	1.90
		WM	0.09	(0.01)	0.00	0.23	1.06	(0.01)	0.00	1.21
		2SM	0.18	(0.02)	0.05	0.95	1.09	(0.04)	0.10	2.22
X	None	None	0.38	(0.03)	0.06	0.98	1.02	(0.03)	0.05	1.64
	SL PS	PM	0.23	(0.02)	0.03	0.70	1.01	(0.03)	0.06	1.72
		WM	0.13	(0.01)	0.01	0.43	1.05	(0.03)	0.06	1.71
		2SM	0.12	(0.01)	0.01	0.43	1.00	(0.03)	0.05	1.65
	RI PS	PM	0.23	(0.02)	0.03	0.72	1.01	(0.03)	0.07	1.75
		WM	0.12	(0.01)	0.01	0.38	1.05	(0.03)	0.05	1.72
		2SM	0.12	(0.01)	0.01	0.41	1.04	(0.03)	0.04	1.61
	RIS PS	PM	0.14	(0.01)	0.01	0.43	1.05	(0.04)	0.08	1.86
		WM	0.11	(0.01)	0.01	0.38	1.05	(0.03)	0.05	1.71
		2SM	0.12	(0.01)	0.01	0.39	1.06	(0.04)	0.05	1.67
Z	None	None	0.29	(0.03)	0.04	0.80	1.03	(0.04)	0.05	1.68
	SL PS	PM	0.18	(0.02)	0.02	0.58	1.02	(0.04)	0.07	1.77
		WM	0.12	(0.01)	0.01	0.39	1.06	(0.04)	0.07	1.83
		2SM	0.12	(0.01)	0.01	0.44	1.03	(0.04)	0.05	1.69
	RI PS	PM	0.19	(0.02)	0.02	0.60	1.02	(0.04)	0.07	1.78
		WM	0.12	(0.01)	0.01	0.40	1.06	(0.04)	0.07	1.86
		2SM	0.14	(0.02)	0.01	0.47	1.05	(0.04)	0.06	1.77
	RIS PS	PM	0.13	(0.01)	0.01	0.44	1.05	(0.04)	0.07	1.84
		WM	0.12	(0.01)	0.01	0.40	1.05	(0.04)	0.06	1.80
		2SM	0.14	(0.01)	0.01	0.53	1.06	(0.04)	0.05	1.74
U	None	None	0.12	(0.01)	0.01	0.39	1.02	(0.03)	0.05	1.65
	SL PS	PM	0.14	(0.02)	0.01	0.45	1.03	(0.03)	0.07	1.80
		WM	0.14	(0.02)	0.01	0.48	1.04	(0.03)	0.07	1.81
		2SM	0.15	(0.02)	0.01	0.52	1.01	(0.03)	0.06	1.73
	RI PS	PM	0.14	(0.02)	0.01	0.45	1.03	(0.03)	0.07	1.78
		WM	0.14	(0.02)	0.01	0.47	1.04	(0.04)	0.08	1.85
		2SM	0.15	(0.02)	0.01	0.53	1.01	(0.04)	0.06	1.77
	RIS PS	PM	0.14	(0.02)	0.01	0.46	1.03	(0.03)	0.07	1.82
		WM	0.14	(0.02)	0.01	0.48	1.04	(0.04)	0.08	1.89
		2SM	0.15	(0.02)	0.02	0.56	1.01	(0.03)	0.06	1.79

Table A.5. Design phase simulation study results summarized across replications: Level-1 covariate balance under treatment assignment with level-1 unobserved covariate (L1UN)

	PS	Match	Within-Site ASB				Within-Site VR			
CV	Model	Method	Mean	(MCSD)	Site Var	Site Max	Mean	(MCSD)	Site Var	Site Max
PS	SL PS	None	0.45	(0.04)	0.07	1.11	1.03	(0.03)	0.06	1.71
		PM	0.24	(0.02)	0.03	0.73	1.01	(0.04)	0.07	1.75
		WM	0.09	(0.01)	0.00	0.23	1.05	(0.01)	0.00	1.20
		2SM	0.08	(0.01)	0.01	0.49	0.97	(0.01)	0.01	1.18
	RI PS	None	0.45	(0.04)	0.07	1.12	1.03	(0.03)	0.06	1.71
		PM	0.26	(0.03)	0.03	0.75	1.01	(0.04)	0.07	1.76
		WM	0.09	(0.01)	0.00	0.23	1.06	(0.01)	0.00	1.20
		2SM	0.23	(0.06)	0.21	2.15	0.94	(0.04)	0.04	1.32
	RIS PS	None	0.46	(0.04)	0.07	1.12	1.03	(0.03)	0.06	1.72
		PM	0.15	(0.02)	0.01	0.43	1.07	(0.04)	0.08	1.87
		WM	0.09	(0.01)	0.00	0.23	1.06	(0.01)	0.00	1.19
		2SM	0.22	(0.03)	0.09	1.33	1.04	(0.04)	0.07	1.87
X	None	None	0.36	(0.03)	0.05	0.93	1.03	(0.03)	0.05	1.68
	SL PS	PM	0.21	(0.02)	0.03	0.67	1.02	(0.04)	0.07	1.74
		WM	0.14	(0.01)	0.01	0.45	1.06	(0.04)	0.06	1.78
		2SM	0.12	(0.01)	0.01	0.49	1.01	(0.04)	0.05	1.67
	RI PS	PM	0.22	(0.02)	0.03	0.69	1.02	(0.04)	0.07	1.78
		WM	0.12	(0.01)	0.01	0.39	1.06	(0.04)	0.06	1.75
		2SM	0.12	(0.01)	0.01	0.44	1.05	(0.03)	0.04	1.64
	RIS PS	PM	0.14	(0.01)	0.01	0.44	1.05	(0.04)	0.07	1.85
		WM	0.12	(0.01)	0.01	0.39	1.06	(0.03)	0.05	1.72
		2SM	0.12	(0.01)	0.01	0.40	1.07	(0.03)	0.05	1.69
Z	None	None	0.28	(0.03)	0.03	0.74	1.02	(0.03)	0.05	1.66
	SL PS	PM	0.17	(0.02)	0.02	0.56	1.02	(0.03)	0.07	1.73
		WM	0.11	(0.01)	0.01	0.38	1.05	(0.04)	0.06	1.77
		2SM	0.12	(0.01)	0.01	0.50	1.02	(0.04)	0.05	1.67
	RI PS	PM	0.18	(0.02)	0.02	0.57	1.02	(0.04)	0.07	1.78
		WM	0.13	(0.01)	0.01	0.43	1.05	(0.04)	0.07	1.82
		2SM	0.14	(0.02)	0.01	0.51	1.05	(0.04)	0.06	1.73
	RIS PS	PM	0.14	(0.01)	0.01	0.43	1.04	(0.04)	0.07	1.81
		WM	0.12	(0.01)	0.01	0.43	1.05	(0.03)	0.07	1.79
		2SM	0.14	(0.02)	0.01	0.56	1.05	(0.03)	0.05	1.71
U	None	None	0.36	(0.03)	0.05	0.93	1.03	(0.04)	0.05	1.67
	SL PS	PM	0.38	(0.04)	0.06	0.99	1.03	(0.04)	0.07	1.76
		WM	0.39	(0.03)	0.06	1.06	1.03	(0.05)	0.08	1.88
		2SM	0.41	(0.04)	0.08	1.17	1.02	(0.04)	0.06	1.74
	RI PS	PM	0.38	(0.03)	0.06	1.00	1.04	(0.04)	0.07	1.82
		WM	0.39	(0.03)	0.06	1.06	1.04	(0.05)	0.08	1.88
		2SM	0.41	(0.04)	0.07	1.13	1.02	(0.04)	0.06	1.77
	RIS PS	PM	0.39	(0.04)	0.06	1.04	1.04	(0.04)	0.07	1.82
		WM	0.39	(0.03)	0.06	1.07	1.04	(0.04)	0.08	1.90
		2SM	0.41	(0.04)	0.07	1.10	1.02	(0.04)	0.06	1.74

Table A.6. Design phase simulation study results summarized across replications: Level-2 covariate balance for observed site-level covariate (S)

Assign. Mech.	Pscore Model	Match Method	"S" ASB Mean	"S" ASB (MCSD)	"S" VR Mean	"S" VR (MCSD)
RA	None	None	0.02	(0.01)	1.00	(0.01)
	SL PS	PM	0.02	(0.01)	1.01	(0.01)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.00)	1.00	(0.00)
	None	None	0.02	(0.01)	1.00	(0.01)
	RI PS	PM	0.02	(0.01)	1.01	(0.01)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.00)	1.00	(0.00)
	None	None	0.02	(0.01)	1.00	(0.01)
	RIS PS	PM	0.01	(0.01)	1.01	(0.01)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.00)	0.99	(0.00)
L1OB	None	None	0.14	(0.03)	1.01	(0.03)
	SL PS	PM	0.03	(0.01)	1.01	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.00)	1.00	(0.01)
	RI PS	PM	0.02	(0.01)	1.01	(0.01)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.00)	1.00	(0.01)
	RIS PS	PM	0.02	(0.01)	1.01	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.00)	1.00	(0.01)
L2OB	None	None	0.50	(0.05)	1.12	(0.06)
	SL PS	PM	0.10	(0.02)	1.04	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.01	(0.01)	0.95	(0.02)
	RI PS	PM	0.08	(0.02)	1.01	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.01	(0.01)	0.97	(0.02)
	RIS PS	PM	0.08	(0.01)	1.10	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.01	(0.01)	1.03	(0.02)

Continued on next page …

Table A.6 continued. Design phase simulation study results summarized across replications: Level-2 covariate balance for observed site-level covariate (S)

Assign. Mech.	Pscore Model	Match Method	"S" ASB Mean	"S" ASB (MCSD)	"S" VR Mean	"S" VR (MCSD)
L2UN	None	None	0.48	(0.06)	1.11	(0.07)
	SL PS	PM	0.10	(0.02)	1.02	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.02)	0.94	(0.03)
	RI PS	PM	0.07	(0.02)	0.98	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.01)	0.96	(0.05)
	RIS PS	PM	0.06	(0.02)	1.07	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.01)	1.03	(0.05)
L1UN	None	None	0.49	(0.06)	1.13	(0.07)
	SL PS	PM	0.10	(0.02)	1.04	(0.05)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.02)	0.94	(0.04)
	RI PS	PM	0.07	(0.02)	0.99	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.02)	0.96	(0.06)
	RIS PS	PM	0.06	(0.02)	1.06	(0.02)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.01)	1.03	(0.06)

Table A.7. Design phase simulation study results summarized across replications: Level-2 covariate balance for unobserved site-level covariate (V)

Assign. Mech.	Pscore Model	Match Method	"V" ASB Mean	"V" ASB (MCSD)	"V" VR Mean	"V" VR (MCSD)
RA	None	None	0.01	(0.03)	1.00	(0.03)
	SL PS	PM	0.02	(0.03)	1.00	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.01)	1.00	(0.01)
	None	None	0.01	(0.03)	1.00	(0.03)
	RI PS	PM	0.02	(0.03)	1.00	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.01)	1.00	(0.01)
	None	None	0.01	(0.03)	1.00	(0.03)
	RIS PS	PM	0.02	(0.03)	1.00	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.00	(0.01)	1.00	(0.01)
L1OB	None	None	0.12	(0.05)	1.00	(0.05)
	SL PS	PM	0.03	(0.04)	1.01	(0.04)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.01	(0.01)	0.99	(0.01)
	RI PS	PM	0.08	(0.03)	1.00	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.01	(0.01)	1.00	(0.02)
	RIS PS	PM	0.08	(0.03)	1.00	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.01	(0.01)	1.00	(0.02)
L2OB	None	None	0.13	(0.18)	1.02	(0.12)
	SL PS	PM	0.03	(0.04)	1.00	(0.05)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.05)	0.99	(0.03)
	RI PS	PM	0.07	(0.03)	1.00	(0.04)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.03)	1.01	(0.03)
	RIS PS	PM	0.07	(0.04)	1.00	(0.04)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.02	(0.03)	1.01	(0.03)

Continued on next page …

Table A.7 continued. Design phase simulation study results summarized across replications: Level-2 covariate balance for unobserved site-level covariate (V)

Assign. Mech.	Pscore Model	Match Method	"V" ASB Mean	"V" ASB (MCSD)	"V" VR Mean	"V" VR (MCSD)
L2UN	None	None	0.28	(0.18)	1.10	(0.15)
	SL PS	PM	0.21	(0.05)	1.09	(0.09)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.12	(0.06)	1.04	(0.06)
	RI PS	PM	0.09	(0.03)	1.01	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.17	(0.05)	1.07	(0.10)
	RIS PS	PM	0.07	(0.04)	1.04	(0.04)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.18	(0.06)	1.08	(0.11)
L1UN	None	None	0.34	(0.17)	1.15	(0.17)
	SL PS	PM	0.28	(0.05)	1.15	(0.13)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.17	(0.06)	1.08	(0.08)
	RI PS	PM	0.09	(0.03)	1.01	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.22	(0.05)	1.11	(0.12)
	RIS PS	PM	0.08	(0.03)	1.04	(0.03)
		WM	0.00	(0.00)	1.00	(0.00)
		2SM	0.23	(0.07)	1.11	(0.14)

Table A.8. Design phase simulation study results summarized across replications: Effect estimation bias

Assign.	Pscore	Match	Grand-Mean ATT					Site-Level Effect Variance		
Mech.	Model	Method	Bias	(MCSD)	Std. Err.	RMSE	Cover.	Bias	(MCSD)	RMSE
RA	None	None	0.00	(0.07)	0.06	0.07	0.91	0.02	(0.03)	0.03
	SL PS	PM	0.00	(0.07)	0.06	0.07	0.92	0.01	(0.03)	0.03
		WM	0.00	(0.07)	0.06	0.07	0.92	0.01	(0.03)	0.03
		2SM	0.00	(0.07)	0.06	0.07	0.91	0.01	(0.03)	0.03
	RI PS	PM	0.00	(0.07)	0.06	0.07	0.93	0.01	(0.03)	0.03
		WM	0.00	(0.07)	0.06	0.07	0.89	0.01	(0.03)	0.03
		2SM	0.00	(0.07)	0.06	0.07	0.90	0.01	(0.03)	0.03
	RIS PS	PM	0.00	(0.07)	0.06	0.06	0.92	0.01	(0.03)	0.03
		WM	0.00	(0.07)	0.06	0.07	0.91	0.01	(0.03)	0.03
		2SM	0.00	(0.07)	0.06	0.07	0.92	0.01	(0.03)	0.03
L1OB	None	None	0.17	(0.07)	0.07	0.18	0.34	0.07	(0.04)	0.08
	SL PS	PM	0.04	(0.06)	0.06	0.07	0.88	0.03	(0.03)	0.04
		WM	0.01	(0.06)	0.06	0.06	0.93	0.01	(0.03)	0.03
		2SM	0.02	(0.06)	0.06	0.06	0.91	0.00	(0.03)	0.03
	RI PS	PM	0.03	(0.06)	0.06	0.07	0.92	0.03	(0.03)	0.04
		WM	0.01	(0.06)	0.06	0.06	0.93	0.01	(0.03)	0.03
		2SM	0.03	(0.06)	0.06	0.07	0.92	0.01	(0.03)	0.03
	RIS PS	PM	0.03	(0.06)	0.06	0.07	0.91	0.02	(0.03)	0.04
		WM	0.01	(0.06)	0.06	0.06	0.92	0.01	(0.03)	0.03
		2SM	0.03	(0.06)	0.06	0.07	0.92	0.01	(0.03)	0.03
L2OB	None	None	0.19	(0.07)	0.08	0.20	0.39	0.13	(0.04)	0.14
	SL PS	PM	0.03	(0.06)	0.07	0.07	0.91	0.04	(0.03)	0.06
		WM	0.02	(0.06)	0.06	0.06	0.90	-0.01	(0.03)	0.03
		2SM	0.01	(0.06)	0.06	0.06	0.90	-0.04	(0.03)	0.05
	RI PS	PM	0.03	(0.06)	0.07	0.07	0.91	0.04	(0.04)	0.06
		WM	0.02	(0.06)	0.06	0.06	0.91	-0.01	(0.03)	0.03
		2SM	0.04	(0.06)	0.06	0.07	0.88	0.00	(0.03)	0.03
	RIS PS	PM	0.01	(0.06)	0.06	0.06	0.90	-0.01	(0.03)	0.03
		WM	0.01	(0.06)	0.06	0.06	0.90	-0.01	(0.03)	0.03
		2SM	0.05	(0.06)	0.06	0.08	0.87	0.01	(0.03)	0.03

Continued on next page …

Table A.8 continued. Design phase simulation study results summarized across replications: Effect estimation bias

Assign.	Pscore	Match	Grand-Mean ATT					Site-Level Effect Variance		
Mech.	Model	Method	Bias	(MCSD)	Std. Err.	RMSE	Cover.	Bias	(MCSD)	RMSE
L2UN	None	None	0.20	(0.08)	0.09	0.22	0.38	0.17	(0.04)	0.17
	SL PS	PM	0.04	(0.06)	0.07	0.07	0.91	0.06	(0.03)	0.07
		WM	0.02	(0.06)	0.06	0.06	0.92	-0.01	(0.03)	0.03
		2SM	0.01	(0.06)	0.06	0.06	0.90	-0.05	(0.03)	0.06
	RI PS	PM	0.04	(0.06)	0.07	0.07	0.90	0.06	(0.04)	0.07
		WM	0.01	(0.06)	0.06	0.06	0.91	-0.01	(0.03)	0.03
		2SM	0.04	(0.06)	0.06	0.07	0.88	0.00	(0.03)	0.03
	RIS PS	PM	0.01	(0.06)	0.06	0.06	0.92	-0.02	(0.03)	0.03
		WM	0.01	(0.06)	0.06	0.06	0.92	-0.01	(0.03)	0.03
		2SM	0.06	(0.06)	0.07	0.08	0.86	0.03	(0.03)	0.04
L1UN	None	None	0.27	(0.08)	0.09	0.29	0.14	0.24	(0.04)	0.24
	SL PS	PM	0.12	(0.07)	0.08	0.14	0.73	0.13	(0.04)	0.14
		WM	0.11	(0.07)	0.07	0.13	0.72	0.07	(0.03)	0.08
		2SM	0.10	(0.07)	0.07	0.12	0.68	0.02	(0.04)	0.04
	RI PS	PM	0.13	(0.07)	0.08	0.15	0.67	0.13	(0.04)	0.13
		WM	0.10	(0.07)	0.07	0.12	0.73	0.06	(0.03)	0.07
		2SM	0.13	(0.07)	0.08	0.15	0.61	0.08	(0.04)	0.09
	RIS PS	PM	0.10	(0.07)	0.07	0.12	0.72	0.05	(0.03)	0.06
		WM	0.10	(0.07)	0.07	0.12	0.73	0.06	(0.03)	0.07
		2SM	0.14	(0.07)	0.08	0.16	0.59	0.12	(0.04)	0.12

Table A.9. Analysis phase simulation study results summarized across replications: Effect estimation bias

Assign.	Match	Analysis	Grand-Mean ATT					Site-Level Effect Variance		
Mech.	Method	Model	Bias	(MCSD)	Std. Err.	RMSE	Cover.	Bias	(MCSD)	RMSE
RA	None	SL UN	0.00	(0.07)	0.02	0.07	0.52	na	--	--
		SL CN	0.00	(0.06)	0.01	0.06	0.24	na	--	--
		ML UN	0.00	(0.07)	0.06	0.07	0.91	0.02	(0.03)	0.03
		ML CN1	0.00	(0.06)	0.06	0.06	0.92	0.01	(0.03)	0.03
		ML CN2	0.00	(0.06)	0.03	0.06	0.62	na	--	--
	2SM	SL UN	0.00	(0.07)	0.02	0.07	0.55	na	--	--
		SL CN	0.00	(0.06)	0.01	0.06	0.31	na	--	--
		ML UN	0.00	(0.07)	0.06	0.07	0.92	0.01	(0.03)	0.03
		ML CN1	0.00	(0.06)	0.06	0.06	0.92	0.01	(0.03)	0.03
		ML CN2	0.00	(0.06)	0.03	0.06	0.70	na	--	--
L1OB	None	SL UN	0.32	(0.06)	0.02	0.32	0.00	na	--	--
		SL CN	-0.02	(0.06)	0.01	0.06	0.29	na	--	--
		ML UN	0.17	(0.07)	0.07	0.18	0.34	0.07	(0.04)	0.08
		ML CN1	-0.04	(0.06)	0.06	0.07	0.95	0.00	(0.03)	0.03
		ML CN2	-0.04	(0.06)	0.03	0.07	0.51	na	--	--
	2SM	SL UN	0.07	(0.06)	0.02	0.09	0.31	na	--	--
		SL CN	0.03	(0.06)	0.01	0.06	0.34	na	--	--
		ML UN	0.03	(0.06)	0.06	0.07	0.92	0.01	(0.03)	0.03
		ML CN1	-0.01	(0.06)	0.06	0.06	0.95	0.01	(0.03)	0.03
		ML CN2	-0.01	(0.06)	0.04	0.06	0.79	na	--	--
L2OB	None	SL UN	0.57	(0.07)	0.02	0.57	0.00	na	--	--
		SL CN	0.03	(0.05)	0.01	0.05	0.27	na	--	--
		ML UN	0.19	(0.07)	0.08	0.20	0.39	0.13	(0.04)	0.14
		ML CN1	-0.05	(0.06)	0.06	0.07	0.91	-0.03	(0.03)	0.04
		ML CN2	-0.05	(0.06)	0.03	0.07	0.45	na	--	--
	2SM	SL UN	0.17	(0.06)	0.02	0.18	0.02	na	--	--
		SL CN	0.10	(0.05)	0.01	0.11	0.04	na	--	--
		ML UN	0.05	(0.06)	0.06	0.08	0.87	0.01	(0.03)	0.03
		ML CN1	-0.01	(0.06)	0.06	0.06	0.93	-0.02	(0.03)	0.03
		ML CN2	-0.01	(0.06)	0.03	0.06	0.70	na	--	--

Continued on next page …

Table A.9 continued. Analysis phase simulation study results summarized across replications: Effect estimation bias

Assign.	Match	Analysis	Grand-Mean ATT					Site-Level Effect Variance		
Mech.	Method	Model	Bias	(MCSD)	Std. Err.	RMSE	Cover.	Bias	(MCSD)	RMSE
L2UN	None	SL UN	0.64	(0.07)	0.02	0.64	0.00	na	--	--
		SL CN	0.10	(0.05)	0.01	0.11	0.07	na	--	--
		ML UN	0.20	(0.08)	0.09	0.22	0.38	0.17	(0.04)	0.17
		ML CN1	-0.05	(0.06)	0.05	0.07	0.88	-0.05	(0.03)	0.05
		ML CN2	-0.05	(0.06)	0.03	0.07	0.44	na	--	--
	2SM	SL UN	0.20	(0.05)	0.02	0.21	0.01	na	--	--
		SL CN	0.12	(0.05)	0.01	0.13	0.03	na	--	--
		ML UN	0.06	(0.06)	0.07	0.08	0.86	0.03	(0.03)	0.04
		ML CN1	-0.01	(0.06)	0.06	0.06	0.93	-0.02	(0.03)	0.04
		ML CN2	-0.01	(0.06)	0.04	0.06	0.75	na	--	--
L1UN	None	SL UN	0.72	(0.08)	0.02	0.72	0.00	na	--	--
		SL CN	0.20	(0.05)	0.01	0.21	0.00	na	--	--
		ML UN	0.27	(0.08)	0.09	0.29	0.14	0.24	(0.04)	0.24
		ML CN1	0.05	(0.06)	0.07	0.08	0.86	0.02	(0.03)	0.04
		ML CN2	0.05	(0.06)	0.03	0.08	0.59	na	--	--
	2SM	SL UN	0.33	(0.06)	0.02	0.33	0.00	na	--	--
		SL CN	0.25	(0.06)	0.01	0.26	0.00	na	--	--
		ML UN	0.14	(0.07)	0.08	0.16	0.59	0.12	(0.04)	0.12
		ML CN1	0.09	(0.07)	0.07	0.11	0.76	0.06	(0.03)	0.06
		ML CN2	0.09	(0.07)	0.05	0.11	0.51	na	--	--

Bibliography

Allensworth, E., Nomi, T., Montgomery, N., & Lee, V. E. (2009). College preparatory curriculum for all: academic consequences of requiring algebra and English I for ninth graders in Chicago. *Educational Evaluation and Policy Analysis*, *31*(4), 367–391. doi:10.3102/0162373709343471

Allison, P. D. (2001). *Missing Data*. Quantitative Applications in the Social Sciences. SAGE.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. doi:10.2307/2291629

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion* (1st ed.). Princeton University Press.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. doi:10.1016/j.csda.2010.11.008

Attewell, P., & Domina, T. (2008). Raising the bar: curricular intensity and academic performance. *Educational Evaluation and Policy Analysis*, *30*(1), 51 –71. doi:10.3102/0162373707313409

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. Retrieved from http://CRAN.R-project.org/package=lme4

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, *22*(4), 551–575. doi:10.1002/pam.10154

Burks, L. C. (1994). Ability group level and achievement. *School Community Journal*, *4*(1), 11–24.

Burris, C. C., Heubert, J. P., & Levin, H. M. (2006). Accelerating mathematics achievement using heterogeneous grouping. *American Educational Research Journal*, *43*(1), 137–154. doi:10.3102/00028312043001105

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2012). *The Aftermath of Accelerating Algebra: Evidence from a District Policy Initiative* (No. 69). CALDER Working Paper. Washington, DC: CALDER.

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, *128*(2), pp. 234–266.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313. doi:10.2307/2528036

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

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

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: results from two types of within-study

comparisons. *Multivariate Behavioral Research*, *44*(6), 828–847. doi:10.1080/00273170903333673

Cronbach, L. J. (1976). *Research on Classrooms and Schools: Formulation of Questions, Design and Analysis.* Stanford, CA: Stanford Evaluation Consortium, School of Education, Stanford University. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED135801

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407–424. doi:10.2307/2669377

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.

Fabelo, T., Thompson, M. D., Plotkin, M., Carmichael, D., Marchbanks, M. P., & Booth, E. A. (2011). *Breaking Schools' Rules: A Statewide Study of How School Discipline Relates to Students' Success and Juvenile Justice Involvement*. New York, NY: Council of State Governments Justice Center.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: a statewide experiment. *American Educational Research Journal*, *27*(3), 557 –577. doi:10.3102/00028312027003557

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*(2), 147–194. doi:10.1177/0049124100029002001

Gamoran, A. (1992). The variable effects of high school tracking. *American Sociological Review*, *57*(6), 812–828. doi:10.2307/2096125

Gamoran, A., & Hannigan, E. C. (2000). Algebra for everyone? Benefits of college-preparatory mathematics for students with diverse abilities in early secondary school. *Educational Evaluation and Policy Analysis*, *22*(3), 241–254.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.). Cambridge University Press.

Gitelman, A. I. (2005). Estimating causal effects from multilevel group-allocation data. *Journal of Educational and Behavioral Statistics*, *30*(4), 397 –412. doi:10.3102/10769986030004397

Greenland, S., & Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, *31*(5), 1030 –1037. doi:10.1093/ije/31.5.1030

Hallinan, M. T., & Kubitschek, W. N. (1999). Curriculum differentiation and high school achievement. *Social Psychology of Education*, *3*(1), 41–62.

Hardgrove, L., Godin, D., & Dodd, B. (2008). *College Outcomes Comparisons by AP and Non-AP High School Experiences* (No. 2008-3). New York, NY: The College Board.

Heckman, J. J. (1989). Causal inference and nonrandom samples. *Journal of Educational Statistics*, *14*(2), 159–168. doi:10.2307/1164605

Heckman, J. J., & Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association*, *84*(408), 862–874. doi:10.2307/2290059

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*(1), 69 –88. doi:10.1093/biostatistics/1.1.69

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199 –236. doi:10.1093/pan/mpl013

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14*(3), 205 –227. doi:10.3102/01623737014003205

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

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*(3), 205 –224. doi:10.3102/01623737027003205

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. doi:10.1198/016214506000000447

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, *33*(3), 333 –362. doi:10.3102/1076998607307355

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25*(1), 305–327.

Kaplan, D., & Chen, C. (2011). *Bayesian Propensity Score Analysis: Simulation and Case Study*. Paper presentation presented at the Society for Research on Educational Effectiveness Spring Conference, Washington, DC.

Kelcey, B. (2011a). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis*, *33*(4), 458–482. doi:10.3102/0162373711415262

Kelcey, B. (2011b). *Propensity Score Matching Within versus Across Schools*. Paper presentation presented at the American Educational Research Association Annual Meeting, New Orleans, LA.

Kim, J., & Seltzer, M. (2007). *Causal Inference in Multilevel Settings in Which Selection Processes Vary Across Schools*. CSE Technical Report 708, CRESST/University of California, Los Angeles.

King, G., & Zeng, L. (2005). The dangers of extreme counterfactuals. *Political Analysis*, *14*(2), 131–159. doi:10.1093/pan/mpj004

Leow, C., Marcus, S., Zanutto, E., & Boruch, R. (2004). Effects of advanced course-taking on math and science achievement: addressing selection bias using propensity scores. *American Journal of Evaluation*, *25*(4), 461 –478. doi:10.1177/109821400402500404

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, *21*(1), 121–145. doi:10.1146/annurev.publhealth.21.1.121

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley-Interscience.

Loveless, T. (2008). *The Misplaced Math Student: Lost in Eighth-Grade Algebra* (Brown Center Report on American Education No. 8). Washington, DC: The Brookings Institute. Retrieved from http://www.brookings.edu/reports/2008/0922_education_loveless.aspx

Ma, X. (2005). Early acceleration of students in mathematics: does it promote growth and stability of growth in achievement across mathematical areas? *Contemporary Educational Psychology*, *30*(4), 439–460. doi:10.1016/j.cedpsych.2005.02.001

McNeal, R. B. (1995). Extracurricular activities and high school dropouts. *Sociology of Education*, *68*(1), 62–80. doi:10.2307/2112764

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of Education Data at the Local Level: From Accountability to Instructional Improvement*. US Department of Education. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED511656

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects. *Sociological Methods & Research*, *35*(1), 3 –60. doi:10.1177/0049124106289164

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (1st ed.). Cambridge University Press.

Nomi, T., & Allensworth, E. (2009). "Double-Dose" algebra as an alternative strategy to remediation: effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*(2), 111–148. doi:10.1080/19345740802676739

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206–230. doi:10.3102/0002831207312905

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models : Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks: Sage Publications.

Raudenbush, S. W., & Willms, Jd. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307 –335. doi:10.3102/10769986020004307

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519. doi:i: 10.1162/edfp.2009.4.4.492</p>

Reinisch, J. M., Sanders, S. A., Mortensen, E. L., & Rubin, D. B. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *JAMA: The Journal of the American Medical Association*, *274*(19), 1518 –1525. doi:10.1001/jama.1995.03530190032031

Rickles, J. H. (2011). Using interviews to understand the assignment mechanism in a nonexperimental study. *Evaluation Review*, *35*(5), 490 –522. doi:10.1177/0193841X11428644

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129. doi:10.2307/2291135

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. doi:10.3102/10769986011003207

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R. (2009). *Design of Observational Studies* (1st ed.). New York: Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524. doi:10.2307/2288398

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. doi:10.2307/2683903

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, New Series, *3*(2), 135–146.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581 –592. doi:10.1093/biomet/63.3.581

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. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, *6*(4), 377 –401. doi:10.3102/10769986006004377

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, *47*(4), 1213–1234.

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

Rubin, D. B. (2004a). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, *29*(3), 343–367.

Rubin, D. B. (2004b). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, *31*(2), 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. doi:10.1198/016214504000001880

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

Rubin, D. B., & Thomas, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, *20*(2), 1079–1093.

Rubin, D. B., & Thomas, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, *79*(4), 797 –809. doi:10.1093/biomet/79.4.797

Rumberger, R. W. (1995). Dropping out of middle school: a multilevel analysis of students and schools. *American Educational Research Journal*, *32*(3), 583 –625. doi:10.3102/00028312032003583

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi:10.1037//1082-989X.7.2.147

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, *13*(4), 279–313.

Schmidt, W. H. (2004). A vision for mathematics. *Educational Leadership*, *61*(5), 6.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating Causal Effects Using Experimental and Observational Designs (Report from the Governing Board of the American Educational Research Association Grants Program)*. Washington, DC: American Educational Research Association.

Sekhon, J. S. (2009). Opiates for the matches: matching methods for causal inference. *Annual Review of Political Science*, *12*(1), 487–508. doi:10.1146/annurev.polisci.11.060606.135444

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-Experiments conducted in field settings. *The Sage handbook of quantitative methodology for the social sciences* (pp. 259–280). Sage Publications.

Shadish, W. R. (2010). Campbell and Rubin: a primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, *15*(1), 3–17. doi:10.1037/a0015916

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. doi:10.1198/016214508000000733

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, *60*(1), 607–629. doi:10.1146/annurev.psych.60.110707.163544

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Boston, MA: Houghton-Mifflin.

Smith, J. B. (1996). Does an extra year make any difference? The impact of early access to algebra on long-term gains in mathematics attainment. *Educational Evaluation and Policy Analysis*, *18*(2), 141–153.

Sobel, M. E. (2007). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, *33*(2), 230–251. doi:10.3102/1076998607307239

Steiner, P. M. (2011). *Matching Strategies for Observational Data with Multilevel Structure*. Paper presentation presented at the Society for Research on Educational Effectiveness Spring Meeting, Washington, DC.

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

Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: matching methods for causal inference. *Best Practices in Quantitative Social Science* (pp. 155–176). Thousand Oaks, CA: Sage Publications.

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. doi:10.3102/1076998607306078

Su, Y.-S., & Cortina, J. (2009). What do we gain? Combining propensity score methods and multilevel modeling. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450058

Taylor, L., & Zhou, X. H. (2009). Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *Biometrics*, *65*(1), 88–95. doi:10.1111/j.1541-0420.2008.01023.x

Thoemmes, F. (2009). *The Use of Propensity Scores with Clustered Data: A Simulation Study* (PhD Dissertation). Arizona State University, Arizona, United States. Retrieved from http://proquest.umi.com/pqdlink?did=1895391031&Fmt=7&clientId=1564&RQT=309&VName=PQD

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

Thoemmes, F., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, *46*(3), 514–543. doi:10.1080/00273171.2011.569395

Wang, J., & Goldschmidt, P. (2003). Importance of middle school mathematics on high school students' mathematics achievement. *The Journal of Educational Research*, *97*(1), 3–17. doi:10.1080/00220670309596624

Williams, T., Haertel, E., & Kirst, M. W. (2011). *Improving Middle Grades Math Performance: A Closer Look at District and School Policies and Practices, Course Placements, and Student Outcomes in California*. Mountain View, CA: EdSource.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–706.