Use of Aggregated Covariates in Propensity Score Analysis of Clustered Data

A dissertation
submitted to the faculty of the
University of Minnesota by

Kyle Nickodem

in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
Department of Educational Psychology
College of Education and Human Development
University of Minnesota

Doctoral Committee:
Ernest C. Davenport, Jr., PhD – Advisor
Michael C. Rodriquez, PhD – Chair
Danielle N. Dupuis, PhD
David M. Vock, PhD

June 2020

ACKNOWLEDGMENTS

I first want to thank my advisor, Dr. Ernest Davenport, for supporting and advocating for me from the moment we met. You offered me opportunities to explore my interests and the space to do so in my own way. Yet, you were always gracious with your time and wisdom when I needed guidance and direction. I am also grateful for the chance to manage the ACT/SAT program with you, which served as a valuable reminder that behind the numbers in a dataset are people with their own unique histories, experiences, and perspectives.

To Dr. Michael Rodriguez and every past and present member of the Minnesota Youth Development Research Group, thank you for introducing me to the propensity score rabbit hole along with countless other measurement and statistical quandaries. The collaboration and camaraderie of this group taught me important lessons about the process of research that I never would have learned in a classroom.

My deepest appreciation goes to Dr. Danielle Dupuis and all of my colleagues at the Research Methodology Consulting Center, for building both my confidence and methodological toolkit. The experiences I gained working with you have been invaluable.

I am also grateful to Dr. David Vock, the EdPsych faculty, and the staff in room 250 for your contributions to my education. You have fundamentally changed how I think and view the world. Above all, you remind me that no act of kindness is too small.

My journey from childhood to doctorate could have been derailed at any number of points without the constant love of my parents. Thank you for always supporting my goals.

To my two exuberant and goofy children, you remind me that there is always more to learn and inspire me to keep asking why, how, when, where, and who. Thank you for also understanding that as much as I want to read you another book, sometimes I have to work too.

Lastly, and most of all, thank you to my patient and loving wife, Kerstin. You have been with me through this entire journey and deserve far more praise than I know how to express. I will keep trying to learn though. I am excited to continue growing with you as we move into our next phase of life.

ABSTRACT

Propensity score methods can be used to reduce selection bias and improve causal inferencing with nonrandomized data. However, there is little guidance for implementing a propensity score analysis when treatment exposure is a property of clusters rather than subjects. For example, education policies and practices are often implemented by school or district rather than by individual student. The three studies in this dissertation strive to clarify procedural quandaries for a propensity score analysis with cluster-level treatment exposure and subject-level outcomes. Additionally, omission of a true confounder from a propensity score analysis can bias treatment effect estimation. My dissertation also explores the utility of aggregated covariates as replacements for missing true cluster-level confounders. The first simulation study compared four procedures for generating aggregated covariates. The results highlight that: 1) researchers need to verify the comparability of generated samples to real world contexts; 2) a propensity score analysis with cluster-level treatment exposure requires at least 60 clusters. The second simulation compared covariate balance and treatment effect estimation when appraising treatment exposure by subjects or by clusters and including aggregated covariates of varying quality. Treatment appraisal by subjects outperformed appraisal by clusters under certain conditions. When highly correlated (r = .92 - .98) with the missing true confounders, aggregated covariates were viable replacements. The last study applied the guidance from the simulations to statewide survey data. The investigation found no association between the presence of a school resource officer and students' social-emotional well-being and academic performance. A critical caveat is the results may not generalize to student populations that have historically been targeted by discrimination and school violence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

Introduction

Consider the increased presence of school resource officers (SROs) in American schools over the last 20 years (Musu et al., 2019; Robers et al., 2013). SROs are oft-armed law enforcement agents tasked with deterring misbehavior and promoting a safe and supportive school environment conducive for learning. How might we determine whether the presence of an SRO has an effect on students' perceptions of safety and support?

Randomized controlled trials (RCTs), also known as randomized experiments, are regarded as the gold standard for drawing causal inferences between a treatment and an outcome (What Works Clearinghouse, 2020). RCTs are comprised of two phases: the design phase and the analysis phase (Rubin, 2007, 2008; Shadish et al., 2002). In its most simple form, the design phase of an RCT consists of randomly assigning each subject to a treatment or control condition and then exposing them to their condition. In the analysis phase, the outcome is measured and the treatment effect is estimated as the difference in outcome, on average, between the treatment and control subjects. When repeated, RCTs produce unbiased estimation of the treatment effect. Unbiased estimation is achievable due to the random assignment creating probabilistically equivalent treatment and control groups prior to introducing the treatment. This isolates the causal mechanism of the treatment while minimizing alternative explanations for variation in the outcome. To determine the causal effects of an SRO – the treatment – on students' sense of safety and support – the outcomes – we could conduct an RCT. Two issues complicate the

implementation of this study design: 1) the ethical and logistical feasibility of an RCT and 2) when subjects are nested within clusters.

In the realm of education, including our SRO example, implementing an RCT can often be cost prohibitive, unethical, or lack generalizability to realistic situations and populations (Rubin, 1974, 2007). For another example, attempts to determine the causal effects of student grade retention cannot be ethically accomplished by randomly assigning some students to repeat a grade while promoting their peers. When random assignment is untenable, nonrandomized (e.g., observational, quasi-experimental) designs must be relied upon. Nonrandomized designs, however, are susceptible to selection bias. For instance, subjects may self-select into the treatment and control conditions rather than being randomly assigned. As a result, the researcher can no longer determine if differences between the treatment and control groups on the outcome are due to the unique effect of the treatment or the effect of a pre-existing covariate, such as socioeconomic status, level of education, or another context specific factor.

Propensity Score Methodology

To ameliorate selection bias and improve causal inferencing in nonrandomized studies, Rosenbaum and Rubin (1983) introduced methods utilizing the propensity score (PS) to mimic an RCT. The PS is defined as the probability of treatment assignment conditional on a set of covariates. When the PS is estimated with a correctly specified statistical model, the treatment and control groups are balanced on the covariates included in the model. As with random assignment in an RCT, this signifies the two groups are probabilistically equivalent. The treatment effect can then be estimated while minimizing the possible influence of the covariates. The steps required to conduct a PS

analysis for treatment effect estimation have been outlined thoroughly in numerous journal articles (e.g., Austin, 2011; Caliendo & Kopeinig, 2008; D'Agostino, 1998) as well as a comprehensive text by Guo and Fraser (2014). The procedures are briefly summarized here with deeper discussions presented in the subsequent chapters.

Steps in a Propensity Score Analysis

The first step is selecting the covariates to include in the PS estimation model. Ideally, the estimation model includes all true confounders – covariates related to both treatment exposure and the outcome. For instance, school racial composition is associated with both the presence of security measures (Servoss, 2017) and perceptions of school climate (Voight et al., 2015). Therefore, when estimating the treatment effect of SRO presence on students' sense of safety and support, school racial composition needs to be included in the model. Additionally, covariates should be measured prior to treatment exposure (Kainz, et al., 2017; What Works Clearinghouse, 2020). In RCTs these are commonly called baseline covariates. Many nonrandomized studies, however, are cross-sectional where covariates and treatment exposure are measured at approximately the same point in time. In this case, covariates that are expected to change as a result of treatment exposure – sometimes called endogenous variables – should be excluded from the PS estimation model. One further consideration is the measurement reliability of the covariate. If a covariate is unreliable, including it in the PS model could potentially increase estimation bias rather than reduce it. One aspect of this dissertation is examining the relationship between estimation bias and the psychometric quality of certain covariates.

The second step is estimation of the PS, which requires specifying the PS model and choosing an estimation method. The functional form of the PS model is primarily informed by the nature of the association between treatment exposure and the covariates selected in step one. A possible specification dilemma is whether interaction or quadratic terms should be included in the model. The estimation method is largely dependent on the type of treatment exposure. For instance, the presence of an SRO is a dichotomous treatment (i.e., either an SRO is present or not). Logistic regression is the most commonly used estimation method for dichotomous treatments in social science research (Thoemmes & Kim, 2011). Other options include probit regression or a variety of machine learning approaches such as classification trees, random forests, and generalized boosted modeling (Cham & West, 2016), as well as a relatively new extension of logistic regression called covariate balancing propensity score (Imai & Ratkovic, 2014). For continuous treatments, such as the dosage of a reading intervention, linear regression can be used.

The third step in a PS analysis is conditioning, also referred to as equating, the treatment and control groups on the PS. In their seminal paper, Rosenbaum and Rubin (1983) proposed matching, stratification, and covariate adjustment conditioning methods to create comparable treatment and control groups. Subsequent weighting methods have also grown in popularity (Austin & Stuart, 2015). Thoemmes and Kim's (2011) literature review found matching as the most common method used in applied social science research. Within the matching approach, a researcher can also choose between a myriad of algorithms (Stuart, 2010). The algorithms vary in their effectiveness to create comparable (i.e., balanced) treatment and control groups and reduce selection bias when

estimating the treatment effect (Austin, 2014; Baser, 2006). As an illustration, consider our investigation of the effect of SROs on students' sense of safety and support. The probability of having an SRO (i.e., the PS) is estimated for each school (or each student) in step two. A matching method with a 1:1 algorithm will pair two schools (or students) that have the same PS, but one with an SRO presence and one without. This creates probabilistically similar treatment and control groups from which the unbiased treatment effect can be estimated.

The fourth step in a PS analysis is to evaluate the degree to which the PS model specified in step two and the conditioning method in step three accomplished the goal of balancing the distributions of the covariates between the treatment and control groups. There are a multitude of approaches to evaluating balance. What Works Clearinghouse (2020) advocates for using the absolute standardized mean difference. Simulation studies concur, demonstrating that the absolute standardized mean difference selects the correctly specified PS model more often than other balancing techniques (Ali, et al., 2014; Austin et al., 2007). Additionally, the absolute standardized difference is not influenced by sample size and can be used equivalently for subject- and cluster-level covariates.

Steps one through four are intended to mimic the design phase of an RCT. The fifth and final step is the analysis phase estimating the treatment effect. If covariate balance is established in step four, the treatment effect can be estimated with a method suitable for the research questions in each specific study. A simple mean difference in the outcome between the treatment and control groups can be sufficient for some studies whereas a complex multilevel latent variable model might be required in other studies.

Assumptions

Unbiased estimation of the treatment effect via PS methods requires satisfying three key assumptions (Imbens, 2004; Rosenbaum & Rubin, 1983). First, the overlap (or positivity) assumption asserts that each unit has a non-zero probability of assignment to the treatment or control condition. In practical terms, for all units $0 < PS < 1$ and the distribution of the PS for the treatment and control groups must overlap. Similar to an RCT, satisfying the overlap assumption allows the balance of the treatment and control groups to be compared in a probabilistic, rather than deterministic, manner. Second, the unconfoundedness (or exchangeability) assumption posits that all baseline covariates associated with the outcome that systematically differ between treatment and control groups (i.e., confounders) must be included in the PS estimation model. The random assignment mechanism in an RCT can balance treatment and control groups on both measured and unmeasured covariates. In a PS analysis, however, only measured covariates – those included in the PS estimation model – can be balanced. Thus, if a confounder is unmeasured, estimation of the treatment effect could be biased (Kainz et al., 2017).

The overlap and unconfoundedness assumptions combine to satisfy the strongly ignorable treatment assignment assumption: given the measured baseline covariates, potential outcomes are independent of treatment assignment. In other words, there is no hidden bias affecting the potential outcomes if the treatment assignment process balances the treatment and control groups on the baseline covariates. A distinguishing difference between RCTs and nonrandomized studies is that the strongly ignorable treatment assignment is typically met in an RCT by virtue of the randomization mechanism, but

rarely met in nonrandomized studies. Consequently, the capacity to draw causal inferences from nonrandomized studies is limited. Causal inferences from nonrandomized studies can, however, be strengthened through PS methods. First, PS methods explicitly quantify whether the overlap assumption is met. Additionally, PS methods provide an avenue for reducing the risk of violating the unconfoundedness assumption. One purpose of my dissertation addresses the unconfoundedness assumption. In particular, when a confounder is unmeasured, can a related variable be used as a proxy? If so, how closely associated must the proxy variable be to the unmeasured confounder in order to reduce bias, rather than introduce more noise, in the treatment effect estimation?

The third assumption is concerning for both RCTs and PS analyses – the stable unit treatment value assumption, or SUTVA (Hong & Raudenbush, 2006; Imbens, 2004; Rosenbaum & Rubin, 1983). SUTVA holds if subjects have complete fidelity to the treatment assignment and the outcome of each subject is independent of the outcome of other subjects. In education settings, however, SUTVA can be easily violated. If a student assigned to a new reading curriculum shares the tips and strategies with their peer in the control condition, SUTVA is violated. The outcomes of the two students are no longer independent nor is the control student adherent to their assigned condition.

Clustered Data

To address infidelity to treatment assignment, rather than assign individual subjects to the treatment or control conditions, assignment occurs by clusters of subjects, such as by classroom or school. SUTVA, however, will still be violated if the subject-level outcome is dependent on cluster membership (Bloom et al., 1999; Feller & Gelman,

2015; Raudenbush, 1997). For instance, suppose student test scores vary by school with some schools having, on average, higher scores than other schools. When evaluating the effect of the new reading curriculum, the students' scores are dependent on the school they attend in addition to any differences due to the curriculum used. In other words, scores of students from the same school will be more similar than scores of students from two different schools regardless of whether they learned from the old or new curriculum. When outcomes are dependent upon cluster membership, the standard errors from the treatment effect estimation can be incorrect because SUTVA has been violated.

Multilevel Models

To account for cluster dependency and appropriately adjust standard errors in the treatment effect estimation, multilevel models (e.g., hierarchical linear models, random effects models) can be used in the analysis phase of an RCT or step five of a PS analysis (Feller & Gelman, 2015; Raudenbush & Bryk, 2002). A multilevel model will typically include subject covariates at Level 1 (L1) and cluster covariates at Level 2 (L2). This scenario is the focus of my dissertation; therefore, the terms L1 and subject-level will be used interchangeably as will the terms L2 and cluster-level. That being said, many of the issues discussed and examined can also apply to longitudinal studies with multiple measurements of the same subjects. With longitudinal data, time is represented at L1 with subjects at L2.

To examine the effect of an SRO on students' sense of safety and support, student characteristics are included at L1 with school characteristics at L2. In a RCT, the treatment indicator is at L2 if schools were randomly assigned to have an SRO or not. The treatment indicator is at L1 if random assignment was by student. PS analyses are

defined by the lack of random assignment, so it is unclear whether the treatment should be appraised at L1 or L2. For the model estimating the treatment effect on the subject-level outcome in step five of a PS analysis, this decision is purely conceptual[1]. A multilevel model estimates the treatment effect the same way regardless of whether it is appraised at L1 or L2[2]. The decision, however, has procedural consequences for selecting covariates in step one, estimating the PS in step two, and conditioning on the PS is step three. For example, if the presence or absence of an SRO is a property of schools, then we would estimate the PS by school using only school covariates and a single-level model. Consequently, every student within a school has the same PS. Conditioning on the PS would also occur by school. Conversely, the treatment could be conceptualized as whether a student attends a school with or without an SRO. In this case treatment is a property of students and the PS is estimated by student. The PS model could include both student (L1) and school (L2) covariates. Students within the same school could have a different PS, and therefore, conditioning on the PS would also occur by student.

Most of the research on the use of PS methods with clustered data, especially in educational contexts, has focused on situations where treatment exposure is by subject. There is very little guidance for applied researchers to conduct a PS analysis when treatment exposure is at the cluster level (L2), but the outcomes of interest are at the subject level (L1). Another purpose of my dissertation is to address this gap in the literature.

[1] Although purely conceptual, the decision can have practical implications stemming from ecological or individualistic (atomistic) fallacies.

[2] This assumes the treatment effect is constant (homogenous). Heterogeneity of the treatment effect can be modelled if treatment is appraised at L1 and a random effect for treatment is included at L2 or via interaction terms.

Aggregated Covariates

Regardless of whether treatment exposure is conceptualized as a property of subjects or clusters, inclusion of cluster-level covariates in the PS estimation model is almost certainly necessary in order to satisfy the unconfoundedness assumption. In practice, it can be difficult to collect data on all cluster-level confounders – the covariates associated with both treatment exposure and outcome. When direct measurement of a L2 confounder is unavailable, but measured at L1, one solution is to aggregate the subject data to the cluster level (Raudenbush & Bryk, 2002). Aggregation is typically conducted by calculating the mean of the L1 values within each L2 cluster. For instance, a measure of school climate can be obtained by asking students their individual perception of school climate. The aggregated school-level covariate is then calculated by taking the mean of the student responses within each school. Other summary statistics, such as the median or mode, can also be used to calculate the aggregated value. Additionally, the percentage of subjects above or below a cut point is also a common method for creating aggregated covariates.

In multilevel modeling more generally, not specifically a PS analysis, Schunck (2016) found through a simulation study that using an aggregated covariate produced downwardly biased estimation of the regression coefficient compared to the true L2 covariate. The magnitude of bias decreased as cluster sample size increased, suggesting that the aggregated covariate became a more reliable indicator of the true L2 covariate as the number of subjects within each cluster grew. Nonetheless, even with a cluster size of 80, which was 80% of the population cluster size, the regression coefficient was still substantially biased regardless of the number of clusters.

Use in Propensity Score Analysis

Although commonly used in applied situations, few simulation studies have addressed the role of aggregated covariates in a PS context. Arpino and Mealli (2011) investigated whether a L1 covariate aggregated to L2 and correlated with a missing true L2 confounder could function as a proxy for the missing true L2 confounder when estimating a treatment effect. For example, the true L2 confounder could be median neighborhood income. Correlated with median neighborhood income could be the L1 covariate of receiving free/reduced price lunch with the aggregated L2 proxy being the proportion of students in a school receiving free/reduced price lunch. Compared to a PS model omitting the true L2 confounder without a proxy, Arpino and Mealli (2011) found that including the aggregated L2 covariate reduced the relative bias and mean square error in the estimation of the treatment effect. Their results suggest that aggregated covariates can be used in place of missing true L2 confounders in a PS analysis. This prompts the question: How strongly associated to the missing true L2 confounder does the aggregated L2 covariate need to be in order to improve estimation of the treatment effect rather than introduce noise? Additionally, Arpino and Mealli (2011) did not evaluate covariate balance, a necessary prerequisite for interpretation of the treatment effect estimate. Although there is no a priori reason to expect any difference from non-aggregated covariates, the ability of various PS methods to balance aggregated covariates remains unexamined.

Research Purpose

The purpose of this dissertation is twofold. The first intent is to explore the utility of aggregated covariates for treatment effect estimation in PS analyses. The second is to

provide guidance to applied researchers employing PS methods to analyze clustered data when treatment exposure is at the cluster level, but subject-level outcomes are of interest. My dissertation aims to answer two broad questions:

1) In a PS analysis with clustered data, the treatment effect estimation can be biased if a true cluster confounder is missing. Therefore, how closely related to the missing true cluster confounder must an aggregated covariate be in order to serve as an adequate replacement in the PS analysis?

2) When treatment exposure is at the cluster level, but the outcome of interest is at the subject level, current research is unclear as to whether treatment exposure should be appraised as a subject or cluster property in a PS analysis. Therefore, is covariate balance and treatment effect estimation impacted by appraising treatment exposure at the subject or cluster level in the PS analysis?

To answer these questions, I will conduct three studies – two simulation and one empirical in nature. Each study has a different intended audience. The first study is primarily for research methodologists and explores how aggregated cluster covariates are generated in simulation study designs (Chapter 2). Multiple approaches have been utilized in multilevel modeling and organizational psychology simulation studies, but it is unknown how the choice of data generation process impacts the characteristics of the samples produced. Yet, these characteristics must be known in order to evaluate the efficacy of the simulation results to real world applications.

Whereas the first study is a simulation to inform the design of simulations, the second study is a simulation to directly investigate the utility of aggregated cluster covariates in a PS analysis of clustered data (Chapter 3). In particular, a PS analysis

where treatment exposure is at the cluster level (L2), but subject-level (L1) outcomes are of interest. Although such a design is common in education, few PS simulation studies have broached the topic and many procedural and contextual questions remain unanswered. Thus, the primary audience for the second study is applied researchers interested in estimating the effect of a cluster-level treatment on a subject-level outcome.

The third study will apply the lessons from the second study to empirical data (Chapter 4). Specifically, the study investigates the example highlighted throughout the introduction: To what extent does the presence of an SRO affect students' sense of safety and support? The presence of an SRO in schools is a cluster-level (L2) treatment whereas perceptions of safety and support are subject-level (L1) outcomes. I also examine the effect of SROs on students' commitment to learning and academic performance. The empirical study has two aims and audiences. First, for applied researchers, the study serves as a practical example of how to conduct a PS analysis with cluster-level treatments and subject-level outcomes. Second, for educational policymakers and stakeholders, the results inform the decisions regarding the extent of funding for and role of SROs in schools.

CHAPTER 2

Simulation 1 – Procedures for Simulating Aggregated Covariates

Background

Simulation provides a valuable tool for assessing under controlled conditions how inclusion of aggregated covariates impacts estimation of the treatment effect in a PS analysis. The relevance of the simulation results, however, is contingent on whether the sample datasets generated in the simulation are representative of the data characteristics researchers encounter in practice. Yet, in a systematic review of 677 simulation studies, Harwell et al. (2018) found only 15 (2.2%) checked the adequacy of the simulated data. Previous simulation studies investigating the use of aggregated covariates in non-PS contexts have employed different procedures for generating the sample datasets. From these studies it is unclear how the choice of data generation procedure impacts the characteristics of the resulting samples. Therefore, my first study evaluated the sample characteristics produced from four procedures for generating aggregated covariates. I then investigated if these differences impacted the bias and variance in PS estimation.

The purpose of this study is to inform simulation design. The intent is not to declare one of the four procedures as optimal. Rather, the intent is to better understand the implications of each data generating procedure on the resulting sample characteristics. Consequently, each procedure may have greater utility for different research purposes. For instance, Schunk (2016) examined the use of aggregated covariates in multilevel models whereas Beal and Dawson's (2007) research focused on the use of survey response data. Given the different contexts, these two studies might require different sample characteristics. Understanding how sample characteristics are affected by data

generation procedures is vital for choosing the best simulation design for each research purpose. Yet, the literature review from Harwell and colleagues (2018) suggests that researchers almost uniformly assume the generated samples contain the desired characteristics rather than actually examining their adequacy. In practice, the analysis presented in this study should be incorporated into the standard process for conducting a simulation study, such as the second study presented in Chapter 3. However, I am treating the evaluation of sample characteristics as a separate study for two reasons: 1) to highlight the need for investigating sample adequacy in simulation studies; 2) to compare the impact of four procedures for generating aggregated covariates. In contrast, most simulation studies only need to check if the single selected procedure performed as intended.

The context for the present study is the use of aggregated covariates in a PS analysis of clustered data. In this context, unbiased estimation of a treatment effect requires the measurement of all covariates associated with an outcome that also show pre-treatment differences between treatment and control groups – also called confounders. These include both subject level (L1) and cluster level (L2) covariates. In education settings, L1 covariates might include student demographic characteristics or previous exam scores while L2 covariates could be teacher characteristics or classroom composition. Direct measurement of L2 covariates can sometimes be challenging. When available, the L2 covariate can be derived from aggregating L1 data (Raudenbush & Bryk, 2002). For instance, measurement of school safety might necessitate aggregating students' perceptions of safety to the school level. Of concern to a PS analysis is the extent to which the aggregated L2 covariate is an accurate representation of the missing

true L2 covariate. If the aggregated L2 covariate contains error, estimation of the treatment effect can be biased. A simulation study can evaluate the use of aggregated covariates in a PS analysis of clustered data, but requires the generated samples contain certain characteristics to do so effectively. The sample characteristics of primary interest are: 1) the correlations between the aggregated and true L2 covariates and 2) indicators of the psychometric quality of aggregated covariates.

Quality of Aggregated Covariates

In their summary of covariate selection in the PS literature, Kainz and colleagues (2017) note the utility of a covariate in a properly specified PS model is conditional on having sound psychometric properties. Similarly, whether an aggregated covariate can replace a true L2 covariate in a PS model is likely contingent on its psychometric properties.

Intraclass Correlation. The psychometric quality of an aggregated L2 covariate is commonly evaluated using two indices of the intraclass correlation (ICC; Bliese, 2000; Raudenbush & Bryk, 2002). ICC(1) is a measure of cluster dependency with values ranging from 0 to 1. A value of 0 indicates L1 responses are independent of cluster membership. Higher values signify greater similarity of L1 responses within a cluster. In the multilevel modeling framework, ICC(1) is defined as the proportion of total variation explained by cluster membership. Represented by $ICC(1) = \tau_{00} / (\tau_{00} + \sigma^2)$, τ_{00} is the between-cluster variance and σ^2 is the within-cluster variance. ICC(1) can also be interpreted as a measure of how representative a single L1 response is of the cluster mean at L2. This latter interpretation also hints at the relationship of ICC(1) and cluster size

with ICC(2). ICC(2) is defined as the reliability of cluster means and calculated via the Spearman-Brown formula as:

$$ICC(2) = k*ICC(1) / [1 + (k - 1)*ICC(1)] \quad (1)$$

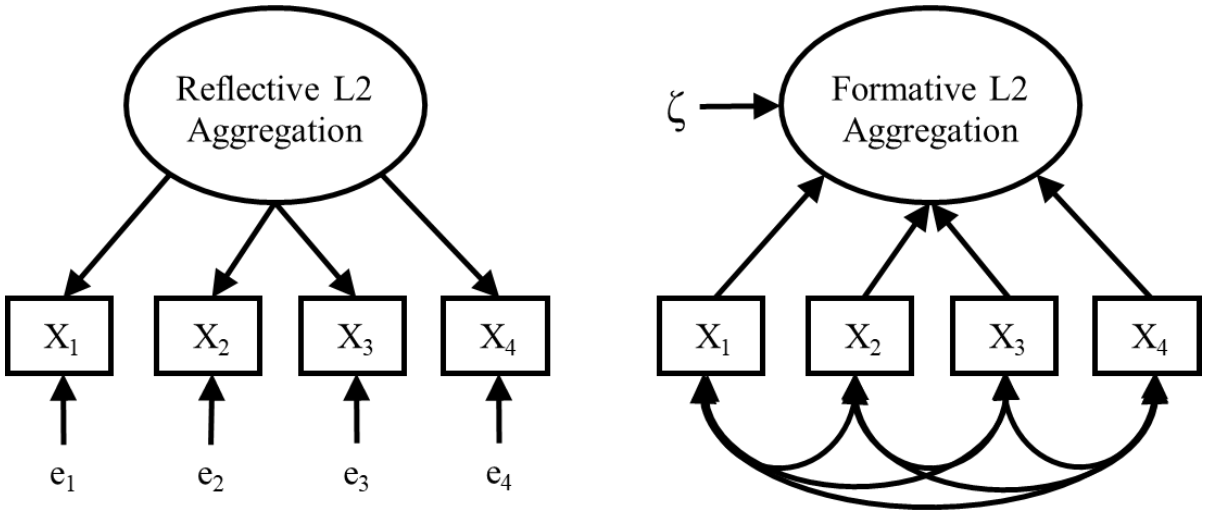
where k is cluster size, or average cluster size when clusters are unbalanced (Bliese, 1998). This formula demonstrates that an aggregated L2 covariate may not be a reliable measure of the true cluster value when cluster size or dependency on cluster membership is small. For example, if cluster size = 20 and ICC(1) = .10 – common values in education – then the reliability of the aggregated L2 covariate values as measured by ICC(2) = .69. One aspect of my dissertation is determining whether a reliability of this magnitude is sufficient for an aggregated L2 covariate to adequately replace a missing true confounder when estimating a treatment effect.

Reflective and Formative Aggregations. Synthezing factor analytic (Kline, 2005) and organizational psychology (Bliese et al., 2007) perspectives on measurement, Lüdtke and colleagues (2008) contend there are different types of aggregated L2 covariates which range on a spectrum from reflective to formative. Furthermore, where an aggregated L2 covariate falls on this spectrum dictates how its psychometric quality should be evaluated. Measures of ICC(1) and ICC(2) as indicators of cluster dependency and the reliability of cluster means, respectively, are intended for use with reflective aggregations. Purely reflective covariates are primarily L2 constructs with the expectation that L1 responses be equally indicative of the true L2 value. In a factor analytic model, reflective aggregrations have the arrows going from the L2 construct to the L1 responses (Figure 2.1). ICC(1) and ICC(2) are expected to be high for reflective

aggregations with 1 – ICC(2) being an indicator of L2 sampling error. Examples include student ratings of teacher ability or measures of school safety.

Figure 2.1

Factor Analytic Representation of Reflective and Formative Aggregations



Note. Level 1 data ($X_1 – X_4$) aggregated to level 2 (L2). Error in the Level 1 covariates is represented by $e_1 – e_4$ while ζ is error in the formative L2 aggregation.

In contrast, a purely formative covariate is primarily a L1 construct with no expectation of isomorphism at L2. In a factor analytic model, the arrows point from the L1 responses to the L2 construct (Figure 2.1). For instance, racial identity at the student level (L1) is a distinctly different construct from the classroom level (L2) measure of percent of White students. In formative aggregations individual observations are not expected to be representative of the mean cluster value. Thus, ICC(1) and ICC(2) are predicted to be low. Instead, they are more appropriately thought of as measures of heterogeneity within a cluster rather than measures of reliability or psychometric quality. For formative aggregations, the sampling ratio – the ratio of L1 responses in the sample to the L1 responses in the population – provides a more informative measure of quality.

When the sampling ratio for a cluster is close to 1.0, the formative aggregation contains little sampling error or random measurement error. That being said, a small sampling ratio is not commensurate with a low quality aggregation. Especially when the cluster size is large, a random sample can produce a formative aggregation with little error even with a small sampling ratio. Thus for practical purposes, it is particularly difficult to evaluate the quality of formative aggregations without also knowing the quality of the data collection process itself.

The distinction between formative and reflective aggregation is also blurry in practice. When obtaining an aggregated measure of school safety, students could be asked about their own sense of safety at school – a more formative approach. They could instead be asked about the general safety of the school – a more reflective approach. Regardless of the approach, both student perceptions and school characteristics are likely to influence responses at both L1 and aggregated to L2. Consequently, the aggregated L2 covariate is neither purely reflective nor purely formative. Nonetheless, when determining the utility of an aggregated L2 covariate in a PS estimation model, consideration for the degree to which the aggregation is reflective or formative informs the interpretation of the metrics used to judge its quality.

Procedures for Generating Aggregated Covariates

Simulation studies are an excellent avenue for evaluating the impact of aggregated covariates in statistical analyses. There is not, however, a uniform method for creating the aggregated covariates in the simulated datasets. For instance, Bliese (1998) inspected the relationship between cluster size, ICC values, and cluster-level correlations with aggregated covariates. To do so, he first generated the true L2 values for x and y

from a normal distribution. The true L2 values were then used to generate the L1 values with the variance varying by simulation condition. The L1 values were then aggregated to create the observed L2 values. The type of aggregation was not explicitly stated in the paper. Given that the L1 values were derived from the true L2 values, this procedure implies the resulting aggregated covariate was reflective in nature. Further evidence of a reflective aggregation came from Bliese's (1998) use of ICC(1) and ICC(2) to evaluate the quality of the aggregated L2 values. Beal and Dawson (2007) followed a similar procedure, but in a very different context. They examined the appropriateness of aggregating a single rating scale item for making cluster-level inferences. The true L2 covariate was generated first from a normal distribution, then used to produce the L1 covariate. Thresholds were then imposed on the continuous L1 covariate to create the rating scale values, which were then aggregated to create the observed L2 values.

Schunk (2016) took the opposite approach and first generated the L1 values when investigating the the effect of small cluster sizes on parameter estimation in multilvel models. Following a formative approach, the L1 values were then aggregated to obtain the true L2 values. A random sample of subjects per cluster were selected with the values from the sample then aggregated to obtain the observed L2 values. In other words, sampling ratio was a manipulated factor in the study rather than ICC(1) or ICC(2). This aligns with the framework for formative aggregations posited by Lüdtke and colleagues (2008).

Present Study

Although all of these papers were interested in exploring the effects of using an aggregated L2 covariate compared to the true L2 covariate, none of the studies first

demonstrated the similarity of the aggregated and true L2 values before examining the effects. The purpose of this simulation is to examine how different procedures for generating the true and aggregated L2 values affects the correlation between them. This is not to say that one procedure is better than the other. Rather, a procedure might yield sample characteristics more aligned with answering certain research questions. For my specific purpose, I am interested in whether the procedures create samples with characteristics desirable for use in a PS analysis of clustered data. Specifically, a PS analysis where treatment exposure is at the cluster level. Therefore, additional sample characteristics of interest are: a) the magnitude of cluster dependency as measured by ICC(1), b) the reliability of the aggregated L2 values as measured by ICC(2), and c) the proportion of subjects in the treatment and control groups. I would also expect other sample characteristics to remain constant based on the simulation design, such as the variance of each L1 and true L2 covariate and the correlation between the covariates within each level. The simulation aims to answer the following research questions:

1. What is the correlation between the true and aggregated L2 values when data are generated from four different simulation procedures?

2. How are the correlations affected by contextual factors of cluster size, number of clusters, ICC(1) of covariates to be aggregated, and magnitude of error in the aggregated L2 values?

3. Do the four data generation procedures produce different sample characteristics, including cluster dependency of aggregated covariates at L1 and the L1 outcome, reliability of aggregated covariates at L2, correlations between covariates at each level, and proportion of subjects in treatment and control groups?

4. What is the bias and precision of PS estimation using the aggregated L2 covariates generated from the four generation procedures compared to the true PSs computed from the true L2 covariates?

Method

Data Generation

For simulation studies to have any practical value, the datasets generated must have realistic sample characteristics. Therefore, it is important for us to consider the characteristics we expect in a dataset used for a PS analysis in the field of education. In a systematic review of 79 studies, albeit none utilizing multilevel models, Thoemmes and Kim (2011) found PS studies included an average of 31 covariates. Therefore, in the present study all datasets contained 20 L1 covariates and 10 L2 covariates along with a dichotomous treatment indicator at L2, Z_j, and outcome variable at L1, Y_{ij}. In education data, covariates are rarely uncorrelated. Thus, the covariates at each level were intended to be correlated at $\rho = .20$, considered a weak to moderate association (Cohen, 1988).

After generating the 20 L1 and 10 true L2 covariates from one of the four covariate generation procedures detailed below, the true PS for each cluster *j* was calculated from a single-level model using the 10 true L2 covariates. This implies treatment exposure occurs at the cluster level, rather than the subject level.

$$\text{logit}(Z_j = 1) = \beta_0 + \beta_1\bar{X}_{1j} + \beta_2\bar{X}_{2j} + \beta_3\bar{X}_{3j} + \beta_4\bar{X}_{4j} + \beta_5\bar{X}_{5j} + \beta_6\bar{X}_{6j} + \beta_7\bar{X}_{7j} + \beta_8\bar{X}_{8j} + \beta_9\bar{X}_{9j} + \beta_{10}\bar{X}_{10j} + u_j \quad (2)$$

The dichotomous treatment exposure, Z_j, was then determined by:

$$(Z_j = 1) \textit{ if } \text{logit}(Z_j = 1) > 0, \textit{ else } (Z_j = 0) \quad (3)$$

The intercept, β_0, was set to -1.0986 to impose a marginal probability of treatment exposure (i.e., mean PS) of .25 and imply a treatment-to-control ratio of 1:3. Treatments are often initially implemented on a small scale with a larger population of control units to draw from. If the treatment group is too small, however, difficulties achieving covariate balance can arise (Lingle, 2009). The regression coefficients β_1 - β_{10} were fixed to 0.50. Previous studies suggest the magnitude of a non-zero coefficient has little bearing on covariate balance or treatment effect estimation (Bellara, 2013; Kelcey, 2009). The cluster residuals, u_j, were drawn from a logit distribution with $\mu = 0$ and $\sigma^2 = \pi^2/3$, the theoretical variance of a logit distribution. Doing so enabled the treatment and control groups created in Equation 3 to have overlapping probability of treatment distributions. A similar procedure was used by Arpino and Mealli (2011) and Leite et al. (2015). The PS was then estimated using Equation 2, but with the true L2 covariates ($\bar{X}_{1j} - \bar{X}_{10j}$) replaced by the aggregated L2 covariates ($\bar{X}'_{1j} - \bar{X}'_{10j}$).

Lastly, the true outcome, Y_{ij}, for each subject i in cluster j was calculated with a two-level random intercept model:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X_{5ij} + \beta_{6j}X_{6ij} + \beta_{7j}X_{7ij} + \quad (4)$$

$$\beta_{8j}X_{8ij} + \beta_{9j}X_{9ij} + \beta_{10j}X_{10ij} + \beta_{11j}X_{11ij} + \beta_{12j}X_{12ij} + \beta_{13j}X_{13ij} + \beta_{14j}X_{14ij} +$$

$$\beta_{15j}X_{15ij} + \beta_{16j}X_{16ij} + \beta_{17j}X_{17ij} + \beta_{18j}X_{18ij} + \beta_{19j}X_{19ij} + \beta_{20j}X_{20ij} + r_{ij}$$

$$r_{ij} \sim N(0, \sigma^2)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{X}_{1j} + \gamma_{02}\bar{X}_{2j} + \gamma_{03}\bar{X}_{3j} + \gamma_{04}\bar{X}_{4j} + \gamma_{05}\bar{X}_{5j} + \gamma_{06}\bar{X}_{6j} + \gamma_{07}\bar{X}_{7j} + \gamma_{08}\bar{X}_{8j} +$$

$$\gamma_{09}\bar{X}_{9j} + \gamma_{0\,10}\bar{X}_{10j} + \gamma_{0\,11}Z_j + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

…

$\beta_{20j} = \gamma_{20\,0}$

The subject level, r_{ij}, and cluster level, u_{0j}, error terms were generated from a normal distribution with $\mu = 0$ and variances of σ^2 and τ_{00}, respectively, which varied depending on the simulation condition. The grand mean intercept, γ_{00}, was set to 0 and the regression coefficients at both the subject ($\gamma_{10} - \gamma_{20\,0}$) and cluster levels ($\gamma_{01}$ - $\gamma_{0\,10}$) were set to 0.50. The treatment effect, $\gamma_{0\,11}$, was also set to 0.50. All simulations were conducted in *R* (v. 3.6.2; R Core Team, 2020) with correlated covariates generated using the *simstudy* package (v. 0.1.15; Goldfeld, 2020). The R script to run the simulation is documented here: https://github.com/knickodem/AggCovsForPS.

Table 2.1

Summary of Simulation 1 Manipulated Factors and Levels

Factor	Levels
Number of clusters	20, 60, 100
Number of subjects per cluster	20, 60, 100
ICC(1) of outcome & aggregated covariates	.05, .10, .20
Covariate generation procedure	Formative-RE, Formative-Sample, Reflective-RE, Reflective-Sample
Aggregated L2 error magnitude	σ of random error: .1, .3, .5, 1 Sampling ratio: 90%, 70%, 50%, 30%

Manipulated Factors

Five factors were manipulated in the simulation (Table 2.1): number of clusters, number of subjects per cluster, ICC(1) of the 10 covariates to be aggregated ($X_{1ij} - X_{10ij}$) and the outcome (Y_{ij}), procedure for generating the aggregated covariates, and magnitude of error imposed between the true and aggregated L2 covariates. The factors were fully

crossed in a 3 (clusters) x 3 (subjects) x 3 (ICC) x 4 (procedure) x 4 (error magnitude) design for a total of 432 conditions. Each condition was replicated 1000 times.

The number of clusters was either 20, 60, or 100. Both the PS and broader multilevel modeling literature suggest that fewer than 50 clusters can lead to estimation problems (Bellara, 2013; Lingle, 2009; Maas & Hox, 2005). Nonetheless, obtaining information on a large number of clusters can be costly in practice and education studies commonly contain fewer than 50 clusters. Likewise, the number of subjects per cluster was 20, 60, or 100. Previous simulation studies have used sizes ranging from 1 (Bellara, 2013) to 200 (Thoemmes & West, 2011) with treatment effect estimation bias decreasing as size increased (Leite et al., 2015; Lingle, 2009). The values in the present study cover the range of a medium-sized classroom to a small school. Educational outcomes typically have ICC(1)s ranging from close to 0 to .25 depending on the construct and whether L2 is at the classroom, school, or district level (Fahle & Reardon, 2018; Nickodem et al., 2019; Servoss, 2017). The present study examined ICC(1)s of .05, .10, and .20.

The covariate data were generated using four different procedures based on two characteristics: 1) mimicking a formative or reflective aggregation, and 2) whether the aggregated L2 covariates were generated by sampling L1 responses (Sample) or adding random error to the true L2 covariate (RE). Table 2.2 summarizes the data generation procedures.

The first procedure is termed the Formative-RE procedure. It imitates a formative aggregation by generating the L1 covariates first with random error added to the true L2 covariates to produce the aggregated L2 covariates. This scenario is akin to using a related, but proxy, L2 covariate to replace a missing true L2 covariate with both the true

and the proxy covariate being formative constructs. For instance, using student free/reduced price lunch status aggregated by school as a proxy for school-level socioeconomic status.

Table 2.2

Summary of Covariate Generation Procedures

Step	Formative-RE	Formative-Sample	Reflective-RE	Reflective-Sample
1	Generate cluster-dependent L1 values	Generate cluster-dependent L1 values	Generate true L2 values (i.e., cluster means)	Generate true L2 values (i.e., cluster means)
2	Calculate true L2 values (i.e., cluster means) from L1 values	Calculate true L2 values (i.e., cluster means) from L1 values	Generate L1 values from true L2 values	Generate L1 values from true L2 values
3	Add random error to true L2 values to create observed aggregate L2 values	Select sample of L1 values from each cluster	Add random error to true L2 values to create observed aggregate L2 values	Select sample of L1 values from each cluster
4		Calculate aggregate L2 values from sample L1 values		Calculate aggregate L2 values from sample L1 values
Ex.	%FRPL in school as proxy for Socioeconomic Status	School racial composition based on student sample	School achievement with measurement error	School climate measured from student sample

Note. RE = Random Error; FRPL = receiving Free/Reduced Priced Lunch

In the Formative-RE procedure, the 20 L1 covariates (X_{1ij}, X_{2ij}, …, X_{20ij}) were generated first from a multivariate normal distribution with $\mu = 0$, variance σ^2, and correlated at $\rho = .20$. To induce cluster dependency, an adjustment for each cluster was drawn from a distribution of $N(0, \tau_{00})$ and added to L1 covariate values. Ten L1 covariates ($X_{1ij} - X_{10ij}$)

were then aggregated within each cluster *j* to obtain the 10 true L2 covariates ($\bar{X}_{1j} - \bar{X}_{10j}$). These were the covariates used to calculate the true probability of treatment (i.e., true PS) and outcome variable, Y_{ij}. For each cluster, random error drawn from $N(0, \sigma^2)$ (σ^2 varied by simulation condition) was added to the true L2 covariates ($\bar{X}_{1j} - \bar{X}_{10j}$) to obtain the aggregated covariates ($\bar{X}'_{1j} - \bar{X}'_{10j}$).

The second procedure, Formative-Sample, followed the same steps as Formative-RE for generating the L1 covariates and determining the true L2 covariates. To obtain the aggregated L2 covariates ($\bar{X}'_{1j} - \bar{X}'_{10j}$), however, a random sample of subjects was selected from each cluster. The sample mean was then calculated. The size of the sample drawn from each cluster varied by simulation condition. By sampling, this procedure represents situations when data from all subjects within a cluster are not available for a formative aggregation.

The third procedure, Reflective-RE, contrasts the previous two procedures by first generating the L2 values. In this manner, the Reflective-RE approach is akin to replacing a missing true L2 covariate with a related L2 covariate. For instance, using a measure of teacher support in place of a measure of school climate. The Reflective-RE procedure began by generating the 10 true L2 covariates ($\bar{X}_{1j} - \bar{X}_{10j}$). An additional 10 L2 covariates were generated solely for the purpose of creating correlated L1 covariates. The values for the 20 L2 covariates were drawn from a multivariate normal distribution with $\mu = 0$, variance τ_{00} (which varied by condition), and correlated at $\rho = .20$. The 20 L1 covariates ($X_{1ij} - X_{20ij}$) were then generated from a normal distribution with a mean equal to the cluster value for the corresponding L2 covariate and variance σ^2, which varied by simulation condition. To obtain the observed aggregate L2 covariates ($\bar{X}'_{1j} - \bar{X}'_{10j}$),

random error was added to the analogous true L2 covariate ($\bar{X}_{1j} - \bar{X}_{10j}$). The random error was drawn from $N(0, \sigma^2)$ with σ^2 varying by simulation condition. In addition to representing a proxy variable replacing a missing confounder, the Reflective-RE procedure also emulates when measurement error in the L1 values is passed on to the aggregated L2 values. For example, statewide achievement test scores are interpreted as reflections of school quality. The observed school achievement scores aggregated from student scores will differ from the (typically unknown) true school achievement scores when student scores contain measurement error.

The fourth and final procedure, Reflective-Sample, followed the same steps as Reflective-RE for generating the true L2 and L1 covariates. The aggregated L2 values ($\bar{X}'_{1j} - \bar{X}'_{10j}$), however, were calculated by selecting a random sample of subjects within each cluster, then aggregating the values to L2. As with Formative-Sample, the size of the sample drawn varied by simulation condition.

The levels manipulated for the magnitude of error between the true and aggregated L2 covariates depended on whether a RE or Sample data generation procedure was used. In the Formative-RE and Reflective-RE procedures, the error was drawn from a normal distribution with $\mu = 0$ and standard deviation of .1, .3, .5, or 1. These levels were chosen arbitrarily to represent aggregated L2 covariates that were increasingly imprecise proxies for the true L2 values. For the Formative-Sample or Reflective-Sample methods, the sampling ratio was 90%, 70%, 50%, or 30% of the subjects within each cluster.

Dependent Variables

To answer the first two research questions, the correlation between each of the 10 true L2 covariates and its aggregated counterpart was calculated, transformed via Fisher's r-to-z formula, and averaged for each replication (Silver & Dunlap, 1987). The mean was calculated across replications for each simulation condition before being transformed back to a correlation with Fisher's z-to-r formula. This provided a measure of how different procedures for generating the aggregated L2 covariates influenced their utility as replacements for the true L2 covariates under various contextual constraints.

For the third research question, various sample characteristics were examined, some of which were expected to vary by simulation condition while others were expected to remain constant. For instance, the ICC(1) for the 10 L1 covariates ($X_{1ij} - X_{10ij}$) that were aggregated and ICC(2) of the resulting aggregated L2 covariates were calculated. Both of these were expected to vary as ICC(1) and the number of subjects were manipulated. Likewise, I expected the variance and ICC(1) of the outcome, Y_{ij}, to vary by simulation condition. The mean variance of the 20 covariates at L1 ($X_{1ij} - X_{20ij}$) was intended to be close to 1.00 and the mean correlation close to $\rho = .20$. The mean variance and mean correlation were also calculated for the 10 true L2 covariates ($\bar{X}_{1j} - \bar{X}_{10j}$) and 10 aggregated L2 covariates ($\bar{X}'_{1j} - \bar{X}'_{10j}$). The mean variance of the L2 covariates was expected to vary by simulation condition, but the correlation between the true L2 covariates should be near $\rho = .20$. With the introduction of random or sampling error into the aggregated L2 covariates, the correlation was expected to be lower than $\rho = .20$.

Lastly, given the context of the simulation being a PS study, the mean true PS (i.e., probability of treatment), mean estimated PS, and proportion of the sample assigned to treatment were calculated for each replication and then averaged across replications.

Additionally, the convergence rate of the PS estimation model was tracked across conditions. For the fourth research question, the bias, mean absolute error (MAE), and root mean square error (RMSE) of the estimated PSs from the true PSs were calculated. Bias provides a measure of systematic error in estimation with positive values indicating overestimation of the PS and negative values indicating underestimation. Bias was calculated within each replication by:

$$\text{Bias} = \frac{\sum_{i=1}^{n}(\widehat{PS}_i - PS_i)}{n} \quad (5)$$

MAE and RMSE both indicate the overall magnitude of error. MAE weights all observations equally whereas RMSE gives greater weight to large errors. MAE and RMSE were calculated within each replication by:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\widehat{PS}_i - PS_i|}{n} \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{PS}_i - PS_i)^2}{n}} \quad (7)$$

Analytic Procedure

The mean for each dependent variable across the 1000 replications was calculated for the 432 conditions. I ran a factorial ANOVA for each dependent variable to determine the amount of variance explained by each of the five manipulated factors (number of clusters, number of subjects per cluster, ICC(1) of aggregated covariates and outcome, generation procedure, and error magnitude) and all two-way interactions. Using the *sjstats* package (v. 0.17.7; Lüdecke 2019), a partial omega-squared (ω^2_P) effect size was calculated for the variance explained by each manipulated factor and interaction. Omega-squared (ω^2) is a less biased estimator of the population effect size than the more common eta-squared (Lakens, 2013). Additionally, unlike ω^2, ω^2_P provides the unique

variation explained by the manipulated factor while accounting for the other factors. This makes ω^2_P more comparable across studies than ω^2. Cohen's (1988) rough interpretation of ω^2_P magnitudes are .01 = small, .06 = medium, and .14 = large. Although, this interpretation was intended for empirical studies and may not be relevant to a simulation study where, by design, the manipulated factors are expected to explain a large portion of the variation in the outcomes. Model assumptions were checked for each factorial ANOVA with a Q-Q plot of the residuals, scatterplot of residuals and predicted values, and calculation of variance inflation factor for each predictor (Williams et al., 2013). Lastly, visual inspection of the variation in the dependent variables by select manipulated factors was facilitated by boxplots or scatterplots with lines of best fit.

Results

For most of the dependent variables, the Q-Q plot and scatterplot of residual and predicted values suggested violations of the normality and homoscedasticity assumptions for the ANOVAs. Consequently, rather than strictly interpret the ω^2_P values, they were used only to help identify the manipulated factor(s) potentially associated with the dependent variable relative to the other factors (see Appendix A). The conclusions drawn from the results instead rely exclusively on descriptive comparisons.

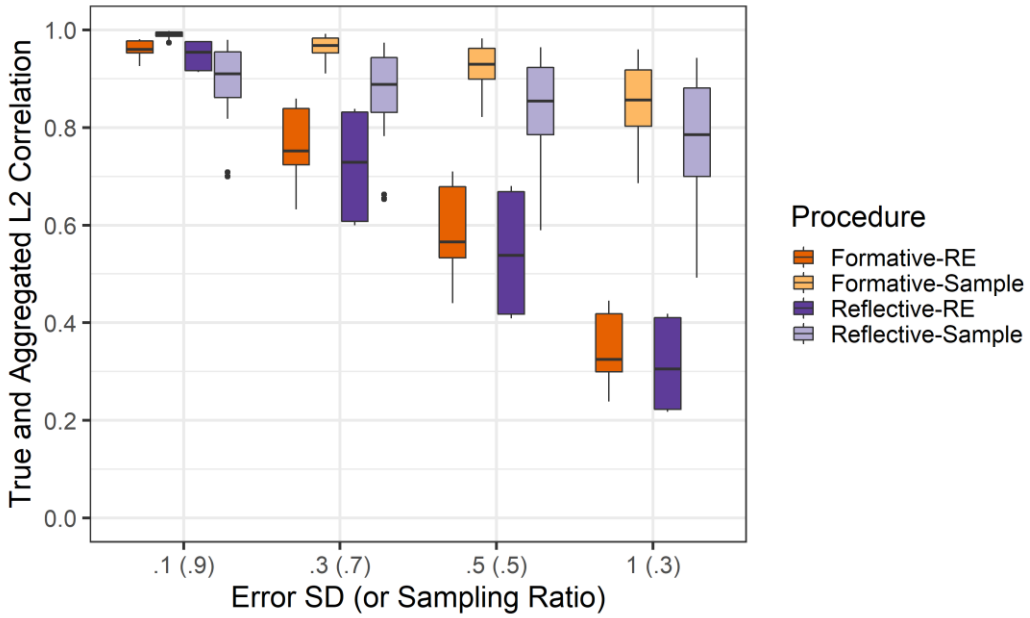
Correlation of True and Aggregated L2 Covariates

The first two research questions asked how the correlation between the 10 true L2 covariates ($\bar{X}_1 - \bar{X}_{10}$) and 10 aggregated L2 covariates ($\bar{X}'_1 - \bar{X}'_{10}$) varied by: 1) data generation procedure and 2) other contextual factors. The simulation results show the correlation was predominantly influenced by the procedure for generating and error magnitude in the aggregated covariates. When the error magnitude was small (90%

sampling ratio or *SD* of the random error = 0.1), the median correlation was > .90 for all generation procedures (Figure 2.2). In general, as error magnitude increased, the median correlation decreased while the variation in correlation increased. Nonetheless, even at the largest error magnitudes, the median correlation was > .75 for the two sampling procedures. Conversely, the random error procedures generated correlations ranging from .22 to .45. Thus, to answer the first research question directly, the correlation between the true and aggregated L2 covariates was influenced to a greater extent by the steps for simulating error (sampling or adding random error) than by taking a formative or reflective approach to generating the L1 and L2 covariates.

Figure 2.2

Mean Correlation Between 10 True and Aggregated Level 2 (L2) Covariates



Note. Error SD is the standard deviation of the random error added to the true covariates to create the aggregated covariates in the RE procedures. Sampling ratio is the proportion of Level 1 values used to create the aggregated covariates in the Sample procedures.

For the second research question, the correlation between the true and aggregated L2 covariates was affected by the magnitude of error in the aggregated covariates. The implication is that when an aggregated covariate contains more error and the correlation with its counterpart true covariate decreases, the efficacy of the aggregated covariate in a PS analysis may also decrease. Other contextual factors influencing the utility of the aggregated covariate were ICC(1) and the number of subjects. Overall, the ICC(1) condition – the measure of cluster dependency of the 10 aggregated covariates at L1 (X_1 – X_{10}) and the L1 outcome (Y) – was positively associated with the correlation between the true and aggregated L2 covariates. As ICC(1) increased from .05 to .10 to .20, the median correlation increased from .76 to .85 to .92. Likewise, as the number of subjects per cluster increased (20, 40, 60), so did the median correlation (.78, .87, .91). These trends suggest the efficacy of an aggregated covariate as a proxy for a missing true L2 covariate in a PS analysis likely improves as the number of subjects and cluster dependency grows.

Other Sample Characteristics

The third research question asked whether the four data generation procedures yield sample datasets with differing characteristics. Of particular interest in the context of a PS analysis with clustered data were a) the cluster dependency of the aggregated covariates at L1 ($X_1 – X_{10}$) and the L1 outcome, Y, as measured by ICC(1), b) the reliability of the aggregated covariates at L2 ($\bar{X}'_1 – \bar{X}'_{10}$) as measured by ICC(2), and c) the proportion of subjects in the treatment and control conditions.

ICC(1) and (2) of Aggregated Covariates. Regardless of the generation procedure, the ICC(1) of the aggregated covariates was equal to the ICC(1) condition with the range of values < .006. Likewise, the ICC(2) of the aggregated covariates

functioned as expected given the ICC(1) and number of subjects per cluster in the condition (Table 2.3). For instance, when ICC(1) = .10 and subjects = 60, the expected ICC(2) = (60*.1) / (1 + (60 – 1)*.1) = .87 (Eq. 1). The generated ICC(2) values were between .85 - .87, indicating high reliability for the aggregated values at L2 in the sample datasets in the ICC(1) = .10 and subjects = 60 conditions.

Table 2.3

Expected and Generated ICC(2) Values

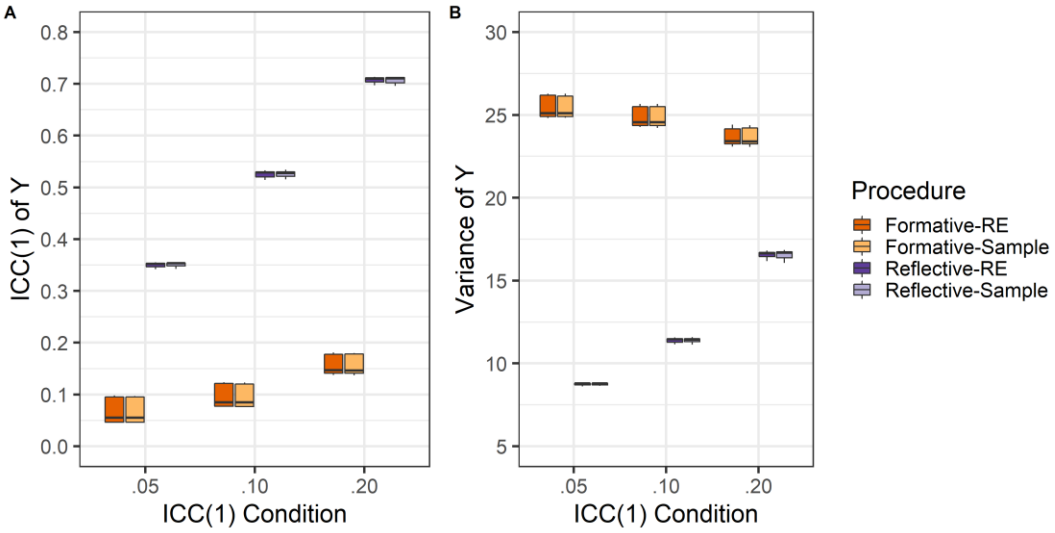
Condition		Expected	Generated ICC(2)	
ICC(1)	# Subjects	ICC(2)	Median	Min - Max
.05	20	.51	.50	.45 - .50
.05	60	.76	.75	.73 - .76
.05	100	.84	.83	.82 - .84
.10	20	.69	.68	.65 - .68
.10	60	.87	.87	.85 - .87
.10	100	.92	.91	.91 - .92
.20	20	.83	.83	.81 - .83
.20	60	.94	.94	.93 - .94
.20	100	.96	.96	.96 - .96

ICC(1) and Variance of *Y*. Although the cluster dependency and reliability for the aggregated covariates did not vary by generation procedure, the same cannot be said about the cluster dependency of the L1 outcome, *Y* (Figure 2.3A). In the ICC(1) = .20 condition, the ICC(1) of *Y* was slightly low for formative procedures (*Mdn* = .15) and extremely high for the reflective procedures (*Mdn* = .71). The latter implies that 71% of the variation in *Y* values generated from the reflective procedures was between cluster variation. The ICC(1) of .20 imposed by the condition is already considered on the high end in education contexts, so a value of .71 is simply unrealistic (Fahle & Reardon, 2018; Nickodem et al., 2019).

Another sample characteristic that varied by generation procedure, and plays a role in the high ICC(1) for reflective procedures, was the variance of *Y* (Figure 2.3B). In the ICC(1) = .05 condition, the variance of *Y* was over twice as large when produced from formative procedures (*Mdn* = 24.5) as from reflective procedures (*Mdn* = 11.4). Additionally, as ICC(1) increased the variance of *Y* decreased in formative procedures, but increased in reflective procedures.

Figure 2.3

Variance and ICC(1) of a Level 1 Outcome (Y) from 432 Simulation Conditions



To understand why, recall that *Y* was created in Equation 4 as the linear combination of the 20 L1 covariates, 10 true L2 covariates, a dichotomous treatment indicator, and random variation at both L1 and L2. We also know that the variance from a linear combination is commonly represented by:

$$Var(aX_1 + bX_2) = a^2Var(X_1) + b^2Var(X_2) + 2abCov(X_1, X_2) \quad (8)$$

In the simulation design, many aspects of Equation 8 were constant across all 432 conditions. All 30 covariates and the treatment indicator (*Z*) had the coefficient $a = b =$

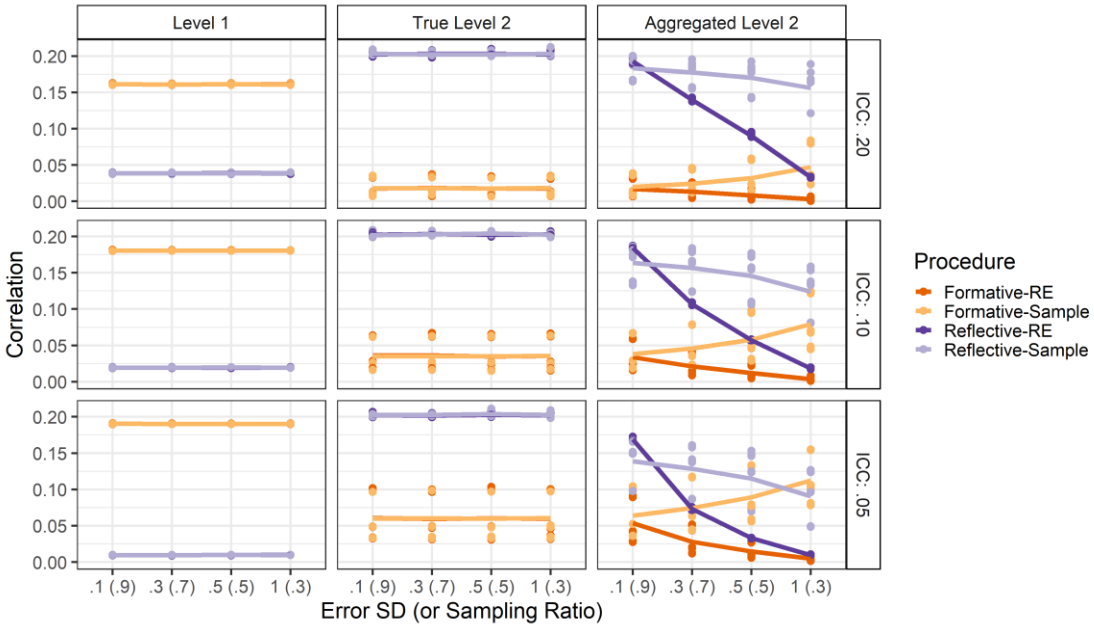
0.50. Additionally, the variance was .99 or 1.00 for all 20 L1 covariates ($X_1 - X_{20}$) for all 432 simulation conditions. The variance of the 10 true L2 covariates ($\bar{X}_1 - \bar{X}_{10}$) showed minimal differences by generation procedure and instead largely aligned with the ICC(1) condition. When ICC(1) = .05, the median variance was 0.05 (range: 0.05 - 0.10); when ICC(1) = .10, the median variance was 0.10 (range: 0.09 - 0.14); and when ICC(1) = .20, the median variance was 0.20 (range: 0.19 - 0.24). Therefore, the variance of the L1 and L2 covariates did not vary substantially by generation procedure. The dissimilarity in the variance of *Y*, and subsequently the ICC(1) of *Y*, by generation procedure must then be due to the last term in Equation 8 – the covariance between the covariates at each level.

Within-level Correlations. In the simulation design, the correlations (i.e., standardized covariances) at each level was set to .20. The correlations within the simulated samples, however, differed depending on the generation procedure (Figure 2.4). The formative procedures generated the L1 covariates first, so it is unsurprising that the observed correlations between the L1 covariates were close to .20 (*Mdn* = .18). When these values were used to create the true L2 values, however, the correlations did not hold (*Mdn* = .03). Conversely, in the reflective procedures, the true L2 covariates were generated first. The imposed correlation of .20 between the true L2 covariates was maintained (*Mdn* = .20), but not for the subsequently created L1 covariates (*Mdn* = .02).

What did this mean for the variance and ICC(1) of *Y*? As the correlation between covariates decreased, the final term in Equation 8 approached 0. The formative procedure maintained the correlation near .20 between L1 covariates, but near 0 for the L2 covariates. The reverse held for the reflective procedures with a correlation near .20 for the L2 covariates, but near 0 for the L1 covariates.

Figure 2.4

Mean Correlation Between Covariates Within Each Level From 432 Simulation Conditions



Consequently, with formative procedures, the variance added to *Y* in the final term in Equation 8 was predominantly from the covariance between L1 covariates. With reflective procedures, however, the final term in Equation 8 was largely constituted by the covariance between L2 covariates. Given that the variance of L1 covariates was greater than the variance of L2 covariates (1.00 > 0.05 – 0.24), but the correlations at each level were similar (~.20), the last term in Equation 8 was larger for formative procedures than reflective procedures. Therefore, formative procedures produced *Y* values with greater variance. Furthermore, with reflective procedures, since the final term in Equation 8 was comprised predominantly of covariance between L2 covariates, a higher proportion of the total variance of *Y* was between cluster variation leading to high

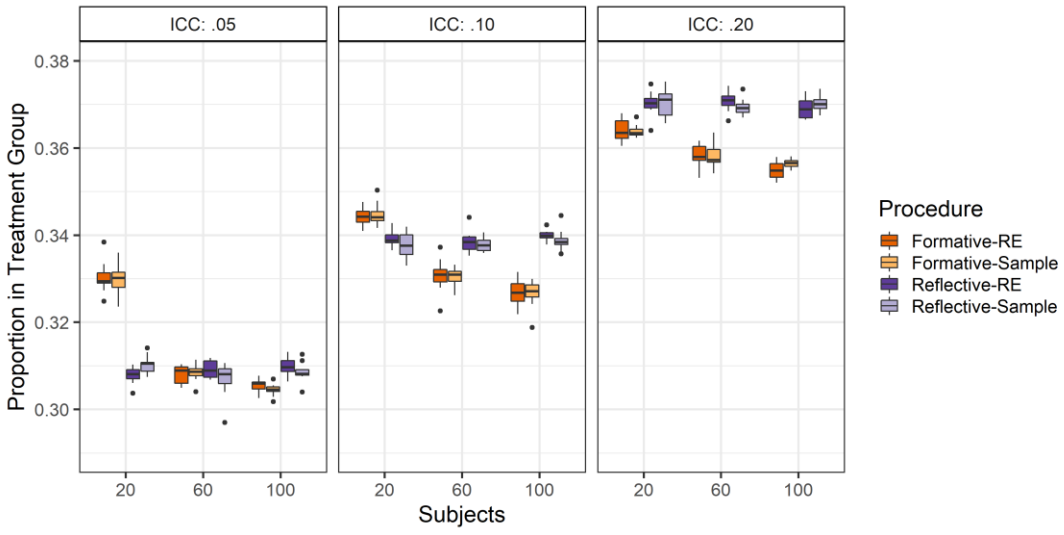
ICC(1) measures. The implications for a PS simulation study are examined in the discussion section.

The correlations between the aggregated L2 covariates are inconsequential for generating *Y* values, but serve as an indication of the generalizability of the simulation results to realistic situations. In practice, covariates are rarely completely independent. Nonetheless, simulation studies often impose independence to simplify the simulation (e.g., Lingle, 2009, Arpino & Mealli, 2011). In the present study, as more random error was added to generate the 10 aggregated L2 covariates, the correlation between them predictably decreased regardless of whether a formative or reflective procedure was used (Figure 2.4). When the sampling ratio decreased, the correlations decreased for the reflective procedure, but actually increased for the formative procedure. Whether this has any implication on the practical application of a simulation study depends on the purpose and context of the study.

Proportion Treated. The last sample characteristic of interest in a PS analysis was the proportion of the sample assigned to the treatment or control group. Imposing an intercept value of -1.0986 in the PS model was intended to produce a marginal probability of treatment exposure of .25. In actuality the proportion across all conditions was a bit higher (*Mdn* = .34) and varied by ICC(1) and the number of subjects in the condition (Figure 2.5). Consequently, the ratio of treatment to control subjects in the sample data was approximately 1:2. Although a deviation from the expected, this is still a very plausible ratio. If a researcher did not check the adequacy of their generated samples, however, they might be interpreting and reporting results under an incorrect assumption about the sample.

Figure 2.5

Proportion of Subjects in the Treatment Group From 432 Simulation Conditions



Propensity Scores Estimation

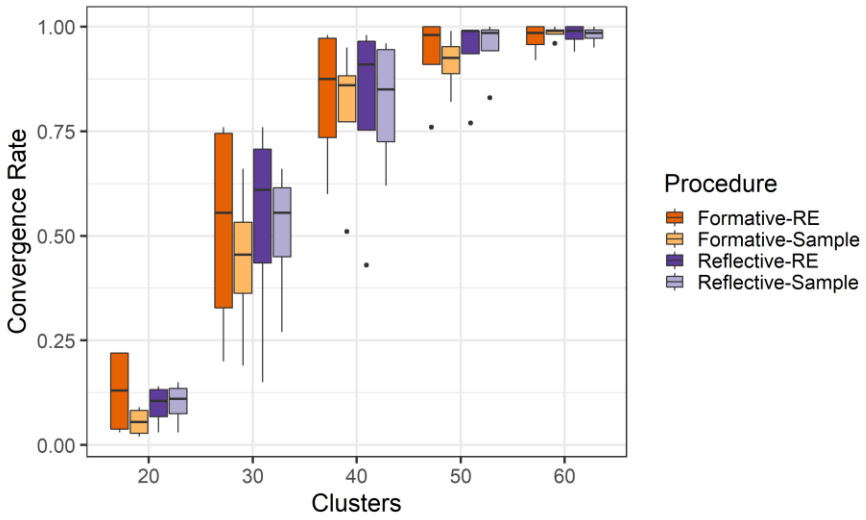
Convergence Rate. The fourth and final research question asked to what extent the use of aggregated covariates impacts bias and variance in PS estimation with treatment exposure at L2 – the cluster level. The first issue to consider is whether or not the model estimating the PS converged. The results indicate the convergence rate was predominately impacted by the number of clusters.[3] With 20 clusters, the median convergence rate for the PS model was only 8.8% (range: 1.2% - 22.8%). The median convergence rate increased to 99.6% (range: 89.9% - 100%) for 60 clusters and 100% (range: 99.7% - 100%) for 100 clusters. In this study, treatment exposure, *Z*, was at L2.

[3] An additional simulation was run to examine whether non-convergence was due to 1) the method for generating the true PS or 2) whether the estimated PS model was too similar to the true PS model despite the substitution of the aggregated covariates for the true covariates. For the former, the method used in the present study and similar to Leite et al., (2015) and others was compared to a method used by Austin, Grootendorst, and Anderson (2007) and a third method similar to sampling dichotomous responses in IRT simulations (e.g. Setoguchi et al., 2008). For the latter, an alternative true PS model included 5 interaction terms. Results showed that non-convergence was still primarily due to the number of clusters.

Therefore, when the PS was estimated with *Z* as the criterion (see Eq. 2), the number of unique PS estimates was equal to the number of clusters. Consider the conditions with 20 clusters and 100 subjects per cluster. Even with a total sample size of 2000, the number of unique PS estimates was still only 20. Consequently, the model overfit the data and produced PS estimates of only 0 and 1, which violates the overlap assumption of a PS analysis. Thus, with a cluster-level (L2) treatment, there needs to be a sufficient number of clusters to make a PS analysis accounting for cluster dependency feasible.

Figure 2.6

Convergence Rate of Propensity Score Model with a Level 2 Treatment by the Number of Clusters and Procedure for Generating Aggregated Covariates



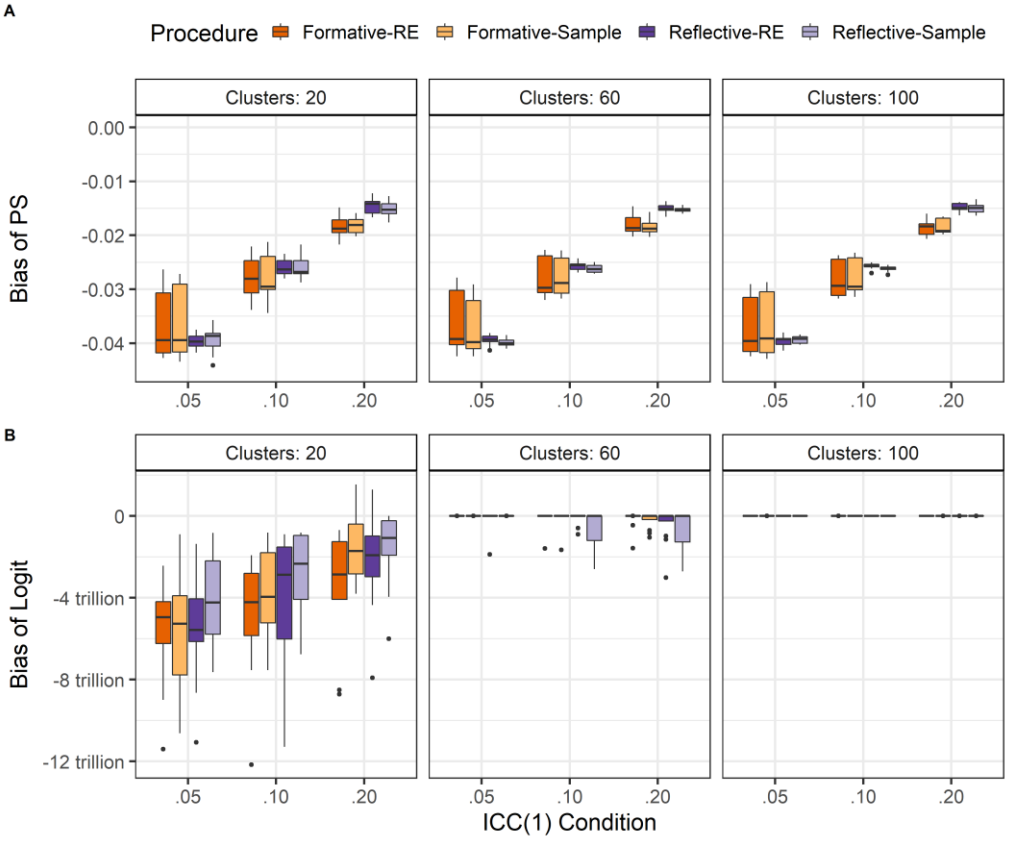
A follow up simulation examined convergence rates for clusters of 20, 30, 40, 50, and 60. The simulation used the same process as the original simulation, but held the number of subjects per cluster constant at 20, and only used the two extreme levels for ICC(1) (.05 and .20) and error magnitude. The results show that with 50 clusters, the median convergence rate was > 98.0% for all procedures except Formative-Sample (*Mdn*

= 92.5%; Figure 2.6). Though even with 60 clusters, the minimum convergence rate was still between 92.0% - 96.0% across the data generation procedures.

Bias. When non-convergence occurs and PS estimates of 0 and 1 are produced, using the PS to examine bias and variance can be misleading. As a probability, the PS is constrained to a range of 0 to 1. The untransformed estimates produced by the PS model are on the logit scale, which has an unconstrained range. Thus, the bias and variance of the estimation in PS units was considerably smaller than the bias and variance of the estimation in logit of the PS units.

Figure 2.7

Estimation Bias with Propensity Score (A) or Logit of the Propensity Score (B) Units



As shown in Figure 2.7A, bias was fairly stable across clusters and largely varied by ICC(1) when calculated from the PSs. When bias was calculated in logit units, however, it becomes clear how the convergence issues with a small number of clusters results in extremely poor estimation (Figure 2.7B). With 60 clusters the issue was largely resolved except for the Reflective-Sample procedure. With 100 clusters, the Formative-Sample (*Mdn* = -0.16) and Reflective-Sample (*Mdn* = -0.10) procedures tended to underestimate the logit of the PS whereas the Formative-RE (*Mdn* = 0.08) and Reflective-RE (*Mdn* = 0.07) procedures overestimated the logit of the PS (Table 2.4). The variation in bias across simulation conditions was also notably larger for Reflective-RE than the other procedures. In application, the bias of 0.08 logits of the PS in the Formative-RE procedure means that if the true PS = .50, the estimated PS, on average, would be .52.

Table 2.4

Median Estimation Accuracy in Logits of the Propensity Score With 100 Clusters

Procedure	Bias	MAE	RMSE
Formative-RE	0.08 (-0.37 - 0.39)	1.78 (1.52 - 2.30)	2.28 (1.95 - 2.93)
Formative-Sample	-0.16 (-0.30 - 0.14)	1.64 (1.49 - 1.89)	2.10 (1.92 - 2.41)
Reflective-RE	0.07 (-5.53 - 0.40)	1.78 (1.52 - 14.76)	2.28 (1.96 - 19.20)
Reflective-Sample	-0.10 (-0.68 - 0.12)	1.65 (1.52 - 3.04)	2.11 (1.96 - 3.84)

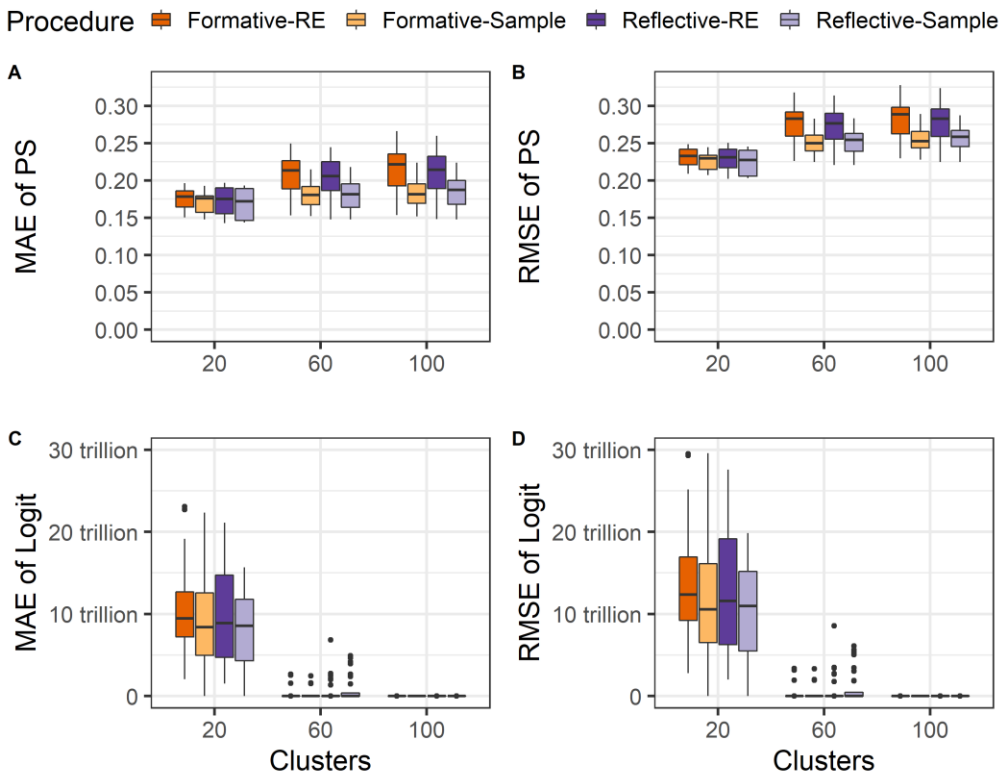
Note. Minimum and maximum values are in parentheses.

MAE and RMSE. As with bias, the measures of MAE and RMSE tell different stories when calculated in PS or logit of the PS units. In PS units, the MAE and RMSE were lower in the 20 cluster conditions than with 60 or 100 conditions (Figure 2.8A and B). In logit of the PS units, however, the MAE and RMSE were in the 10s of trillions for the 20 cluster conditions and substantially smaller for the larger cluster conditions (Figure 8C and D). Thus, calculating estimation error using the PS masks the convergence issues

which are starkly revealed when using the logit of the PS. With 100 clusters and using logits, MAE was slightly lower for the Formative-Sample (*Mdn* = 1.64) and Reflective-Sample (*Mdn* = 1.65) than for random error procedures (*Mdn* = 1.78; Table 2.4). To put the median MAE of 1.64 from Formative-Sample into context, the average distance between the true and the estimated PS was .34, which is larger than the MAE calculated from the PS directly. A similar trend was found with RMSE. Once again, the Reflective-RE procedure had a vastly larger range of MAE and RMSE values than the other procedures, suggesting it performed inconsistently as the other manipulated factors changed.

Figure 2.8

Estimation Mean Absolute Error (A, C) and Root Mean Squared Error (B, D) in Propensity Score (A, B) or Logit of the Propensity Score Units (C, D)



Discussion

The purpose of this study was to examine differences in the sample characteristics and PS estimation resulting from four procedures for generating aggregated covariates in a simulation study. The goal was not to declare one covariate generating procedure as uniformly superior to others, but rather to highlight the consequences of simulation design features on the datasets produced. Harwell et al. (2018) found a woefully low percentage of studies checked, or at least reported checking, the appropriateness of their simulated samples. Failing to do so brings into question the validity of the interpretations and conclusions from the study. Thus, the results here are intended to aid researchers as they decide on the simulation design for generating aggregated covariates and producing datasets appropriate for their particular questions and contexts. The primary conclusions drawn from the present study are: 1) Under the study design constraints, the four procedures for generating aggregated covariates produced samples that differed on a variety of characteristics. Whether the generated characteristics are desired depends on the specific research study. 2) The sample characteristics from the Formative-RE procedure were most suitable for my simulation in Chapter 3. 3) With a cluster-level treatment, at least 60 clusters are needed in order to estimate the PS with minimal non-convergence. 4) When investigating the bias and variance of the PS estimation, use the logit of the PS rather than the PS directly. The truncated scale of the PS can mask estimation issues, especially if convergence warnings are also ignored.

Sample Characteristics

The context of the present study is a PS analysis of clustered data where treatment exposure is at the cluster level (L2). In this context, aggregated covariates can potentially

operate as a proxy for a missing confounder when estimating the PS and the treatment effect. Results from the simulation demonstrate that the procedure for generating aggregated covariates also impacts the characteristics of the sample datasets. Of greatest relevance is the average correlation between each aggregated L2 covariate and its counterpart true L2 covariate. The Formative-Sample and Reflective-Sample procedures that sampled a proportion of the L1 values to aggregate to L2 showed limited change in the correlation even as the sampling ratio decreased to 30%. The lack of variation in the correlation would make it difficult to examine how changes in the correlation between the true and aggregated covariates impacts treatment effect estimation. Conversely, the Formative-RE and Reflective-RE procedures where random error was added to the true L2 values to create the aggregated values produced variation in the correlations sufficient for the task.

Although the data generation procedures that employed sampling yielded datasets unsuitable for my purposes, they might be suitable in other contexts. Schunck (2016) used a procedure similar to Formative-Sample to investigate bias in parameter estimation of multilevel models with aggregated covariates. Schunck found high bias in the parameter estimation even with a sampling ratio of 80%; however, the study did not include the correlation between the true and aggregated L2 covariates. If the pattern aligned with those found in the present simulation, at a sampling ratio of 80%, the correlations would be > .95. The implication is that using an aggregated covariate as a proxy, even a highly correlated one, to replace a missing L2 covariate will increase estimation bias. If Schunk had used the Formative-RE procedure - which produced lower correlations in the present study - would the estimation bias have been even worse?

Another pertinent sample characteristic to a PS analysis with clustered data that varied by covariate generation procedure was the ICC(1) of the L1 outcome, *Y*. If ICC(1) = 0, the data are not clustered; an ICC(1) > .30 is unrealistic in most education contexts. Thus, if the ICC(1) in the simulated samples fell outside this range, the samples were not suitable for the purpose of my study. The Formative-RE and Formative-Sample procedures generated realistic ICC(1)s between .05 and .18. The Reflective-RE and Reflective-Sample procedures, however, did not produce an ICC(1) smaller than .34. Using a Reflective procedure in its current form would clearly be inappropriate for an education centered simulation study. There might be contexts where using a Reflective procedure is more closely aligned with the concepts under investigation, so it is worth considering how changes to the design could improve the efficacy of the procedure. Increasing the variability of the L1 covariates or decreasing the variability of and correlation between the L2 covariates would reduce the ICC(1) of *Y*. That being said, if the researcher did not first check the adequacy of the sample datasets, they would not know whether such adjustments were necessary.

Other characteristics that varied by covariate generation procedure were the correlations between the covariates within each level and the variance of *Y*. Awareness of these characteristic values in the simulated samples allows results from the simulation to be placed in the proper context. For instance, Austin (2009a) determined through a simulation study that as the correlation between the baseline covariates increased, the overlap in the distribution of PS estimates for the treatment and control groups decreased. The implication is that while uncorrelated covariates may not represent realistic situations, high correlations could lead to PS estimation or conditioning issues. Thus, PS

simulation designs need to strike a felicitous balance. I attempted to impose a correlation of .20 between covariates within each level. Using the Formative procedures, the correlation was maintained between the L1 covariates, but not between the true or aggregated L2 covariates. This aligns with findings from Bliese (1998) who noted that the L2 correlations would deteriorate relative to the L1 correlations due to unreliability in the L2 means – which are measured by ICC(2). This consequence from the Formative procedures is acceptable for the purposes in my study. A different research context might deem maintaining the correlation within each level as an essential component. We could ensure the within level correlation by drawing the L1 and L2 covariates from separate distributions. Doing so, however, would make the L2 covariates independent of the L1 covariates rather than aggregations. Accordingly, it is imperative that researchers consider the desired characteristics of the sample datasets when designing a simulation study. Doing so subsequently provides motivation and direction for checking the adequacy of the simulated datasets.

Propensity Score Estimation

In a PS analysis, the accuracy and precision of the PS estimation is typically a tertiary concern following the treatment effect estimation and the covariate balance between the treatment and control groups. Ho and colleagues (2007) note that as a balancing score, as long as the estimated PS enables balancing covariates, the accuracy of the estimated PS to the true PS is immaterial. That being said, the PS model that most closely approximates reality is, theoretically, most likely to produce the best balance (Rubin, 2007, 2008). Of greater concern than approximating reality is when PS estimation is extremely poor. When this occurs, the PSs take values close to 0 and 1,

which not only violates the overlap assumption, but also makes balancing covariates difficult. Thus, checking the adequacy of the PS estimation is a critical step when conducting a simulation study. Results from the present study revealed that evaluating the estimation using the PS directly masks potential problems. Although using the true and estimated PS to calculate bias, MAE, and RMSE is initially more interpretable than using the true and estimated logit of the PS, the results can be misleading due to the constrained probability scale for the PS. Thus, the logit units, which are on an unconstrained scale, provide a more appropriate portrayal of estimation accuracy and precision. This is particularly true when study designs employ a matching algorithm with the logit of the PS as the distance measure to balance covariates (Austin, 2009b).

Number of Clusters. In the present study, poor estimation of the PS resulted primarily from a small number of clusters. Findings suggest that a PS analysis with cluster-level treatment exposure requires > 60 clusters to consistently estimate the PS without convergence issues. This aligns with Pirracchio, and colleagues' (2012) investigation of PS methods with small sample sizes and single-level models. In their simulation a dichotomous treatment exposure with a marginal PS of .20 was estimated with four covariates. Bias and mean square error decreased as the sample size increased from 40 to 1000, but even sample sizes of 40 yielded reasonable estimation of the treatment effect with PS matching or weighting methods. The present study used 10 covariates rather than four, so a larger necessary sample size is expected. As Thoemmes & Kim (2011) found in their systematic review of PS studies utilizing single-level models with subject-level treatments, studies included an average of 31 covariates. If studies show a similar tendency when treatment exposure is at the cluster-level, far more than 60

clusters would be required. Determining the minimum number of clusters needed for PS estimation with a cluster-level treatment under various contextual conditions is a promising line of research. The present study, along with Pirracchio, and colleagues' (2012), utilized logistic regression to estimate the PSs. Probit regression, the covariate balancing propensity score (Imai and Ratkovic, 2014), or machine learning techniques may yield different thresholds for cluster sample size.

Implications for Simulation 2

The simulation study in Chapter 3 investigates the efficacy of aggregated covariates as replacements of missing true covariates for treatment effect estimation of a cluster-level treatment using PS methods. One aim of the present simulation was to inform the design of this subsequent simulation. In addition to illuminating a lower limit for the number of clusters, examination of the sample characteristics produced from the four covariate generation procedures suggest the Formative-RE procedure yields the most suitable sample datasets for my purpose. Compared to the other three procedures, the Formative-RE procedure generated sufficient variability in the correlations between the true and aggregated L2 covariates while producing realistic ICC(1) values for the subject-level outcome, *Y*. Across all generation procedures the proportion of subjects in the treatment group was higher (.30 - .38) than the imposed marginal probability of treatment exposure of .25. Consequently, the ratio of treatment to control subjects was lower than expected. To increase the likelihood of finding adequate matches in the simulation in the next chapter, the marginal probability of treatment exposure will be reduced.

Conclusion

As is the case in applied studies, the study design plays a critical role in the interpretation and conclusions derived from simulations. In particular, the characteristics of the generated sample datasets delineate the real-world scenarios to which the simulation results are most applicable. Unfortunately, the practice of checking the adequacy of the sample datasets is frequently omitted, or at least underreported. Results from the present simulation illustrated that varying the design features and procedures in a simulation can produce samples with starkly different characteristics. Each procedure and set of sample characteristics may have greater utility in different research contexts. Thus, it is the responsibility of the researcher to clarify the characteristics most desired for their research purpose and then examine whether those characteristics are present in the generated sample datasets.

CHAPTER 3

Simulation 2 – Cluster-Level Treatment Exposure and Subject-Level Outcome

Background

When estimating the effect of a treatment on an outcome using PS methods, the specifications of both the PS and outcome models depend on whether treatment exposure and outcome measurement occur on the same or different levels of analysis. In many cases, the treatment and the outcome are both at Level 1 (L1) – the subject level. As an example, Van Boekel and colleagues (2016) conducted a PS analysis investigating the effect of students' participation in sports on their grade point average and perceptions of support. Even with clustered data, the majority of PS research has focused on instances where treatment is administered and the outcome is measured at the same level. In education settings, treatment exposure can also be at Level 2 (L2) – the cluster level – with changes in policy and instruction typically enacted by classroom or school rather than by individual student. Nonetheless, educational stakeholders are often interested in how these broad changes influence outcomes at the student level (L1). Despite the prevalence in education, and relevance to other fields including medical sciences and economics, little methodological consideration has been given in the PS literature to situations where treatment exposure occurs at the cluster level (L2), but outcomes are measured at the subject level (L1). Examples from applied PS studies include the effect of teacher reading knowledge on student reading comprehension (Kelcey, 2011), small school size on student math achievement (Wyse et al., 2008), and college selectivity on students' probability of graduation (Heil et al., 2014). This could also include

longitudinal studies where, for instance, treatment exposure varies by subject (L2) and the outcome of interest is growth over time (L1).

The purpose of the simulation presented in this chapter is to provide guidance for applied researchers conducting PS analyses with cluster-level treatments and subject-level outcomes. In particular, the chapter explores procedural questions: Should the PS be estimated by cluster or by subject? If by subject, how should cluster dependency be addressed in the PS model specification? How many subjects and clusters are needed? What covariates should be included in the PS model? To what extent are aggregated covariates useful as proxies for missing cluster-level covariates in such an analysis? Does the choice of PS conditioning method matter?

Appraisal of Treatment Exposure at the Subject or Cluster Level

In an RCT, the level of the treatment exposure is typically demarcated by the level at which random assignment occurred. What Works Clearinghouse (WWC, 2020) defines cluster-level treatment assignment as situations where: 1) subjects are assigned to treatment and control conditions in groups rather than independently, and 2) outcomes are measured for each subject within a cluster. The subject-level measurements can then be analyzed at the subject level or aggregated and analyzed at the cluster level. A PS analysis is characterized specifically by the lack of random assignment. In this case, WWC (2020) determines the level of treatment exposure by "the largest study unit that contains only members of one condition." (p. 19). One exemplar is Kelcey's (2011) PS analysis with teachers' reading knowledge as the treatment. All students in the classroom of a given teacher were in the same treatment condition. Thus, treatment exposure was appraised at the cluster level in the study. Other studies, however, deviate from the WWC

(2020) standard. For instance, Belfi and colleagues (2016) posed the question, "What are the long-term effects of primary school socio-economic composition on students' mathematics achievement growth?" (p. 503). Socio-economic composition is a property of the school. All students within a school had the same socio-economic composition, i.e. treatment. Yet, the PS was defined in their study as the probability of a student attending a school with a low, medium, high, or mixed socio-economic composition. Thus, the study regarded treatment as a student characteristic.

Other researchers have taken a similar approach to Belfi et al. (2016). This includes studies comparing the effect of large and small schools (Wyse et al., 2008) or charter and traditional public schools (Xiang & Tarasawa, 2015) on student achievement. Both studies defined the PS at the student level (L1) – the probability of a student attending a school with the treatment characteristic. Yet, size and charter designation are properties of the school (L2). Crucially, every student within a school received the same treatment condition. If aligning with the WWC (2020) standard, the PS should instead be defined as the probability of a school having a certain characteristic.

Another perspective to consider is at what level random assignment would have occurred if these studies were designed as RCTs rather than nonrandomized studies. For instance, how might Xiang & Tarasawa (2015) have designed their study comparing student achievement between charter and traditional public schools in an RCT framework? Would they have randomly assigned students or schools to be in the treatment and control conditions? Their study specifically compared students who were in public schools in 5^{th} grade, but some then transferred to charter schools for 6^{th} grade. From this perspective, randomly assigning some students to attend a charter school while

others stay in public schools is a conceivable study design. Although all students within a school received the same treatment, WWC (2020) would not consider this a cluster RCT because random assignment was independent for each subject. Without the clarity of a random assignment mechanism, however, Xiang & Tarasawa (2015) had to decide whether treatment exposure should be appraised by subjects or clusters in the PS model. Their example illustrates the conceptual difficulties applied researchers face when designing a PS study with treatment exposure at the cluster level and measurement of outcomes at the subject level. Furthermore, the decision has procedural consequences for the PS analysis. The decision specifically impacts three steps in a PS analysis: 1) selecting covariates to include in the PS model, 2) specifying the functional form of the PS model, and 3) conditioning on the PS.

Covariate Selection

The first step in a PS analysis is selecting the covariates to include in the PS estimation model. To satisfy the unconfoundedness assumption, the PS model should include all true confounders (Kainz et al., 2017). True confounders are covariates associated with both treatment exposure and the outcome. Ideally, the true confounders are measured prior to treatment exposure (i.e., at baseline). Simulation studies with non-clustered data revealed that omitting a true confounder induced more bias in treatment effect estimation than including a nuisance covariate unrelated to treatment exposure and the outcome (Ali et al., 2014; Austin et al., 2007). Simulations with clustered data produced similar findings when true confounders were excluded at either the subject or cluster level (Kelcey, 2009; Leyrat et al., 2013; Yu, 2012). If a cluster-level treatment exposure is appraised by subject in the PS model, true confounders could be at the subject

or the cluster level. In education settings, subject confounders might include student demographic characteristics or previous exam scores. Cluster confounders could be teacher characteristics or student information aggregated by cluster, such as school composition.

There is currently little research on how covariate selection changes, if at all, when treatment exposure is appraised by clusters in the PS model. On the one hand, there is little expectation for subject covariates to be associated with treatment exposure at the cluster level. For instance, with teachers' reading knowledge as the treatment, there is no a priori reason why student traits would be predictive of teacher knowledge (Kelcey, 2011). Consequently, the PS model would only include cluster-level covariates. On the other hand, the goal of a PS analysis is ultimately to reduce selection bias in the estimation of the treatment effect on the outcome. Research thus far indicates that the PS model should be populated with covariates associated with the outcome even if unrelated to the treatment (Kainz, et al., 2017; Kelcey, 2009). To incorporate subject-level covariates associated with the outcome into a PS model where treatment exposure is appraised at the cluster level requires aggregating the subject information. Kelcey (2011) follows this rationale by including student characteristics (L1; e.g., test scores, demographic variables) aggregated to the teacher/classroom level (L2).

Aggregating covariates can also potentially mitigate the estimation bias induced by the omission of a true cluster confounder from the PS model. Obtaining direct measures of all possible cluster-level true confounders at baseline can be cumbersome or costly. Consider an intervention that relies on groups of students working collaboratively to demonstrate their mathematical abilities. A direct measure of group cooperation prior

to the intervention could be ascertained through raters coding an observation based on a rubric. A more efficient solution, however, could be aggregating the subject information to the cluster level. For instance, the group cooperation measure can be found indirectly by asking students on a survey how cooperative they perceived their group to be and then aggregating the responses. In this example, both direct and aggregated information measured the same construct – group cooperation. In the context of a subject-level treatment, Arpino and Mealli (2011) showed an aggregated covariate measuring a different but correlated construct could be a viable proxy for missing cluster-level true confounder. Their findings suggest that in the mathematics intervention example, if a measure of cooperation is unavailable, a correlated measure could be used instead, such as prosocial behavior (Premo et al., 2018). However, the strength of the association between the aggregated covariate and missing true confounder was unclear in their study. The question becomes, how closely correlated does the aggregated covariate need to be with the missing true cluster confounder in order to reduce bias in treatment effect estimation in a PS study?

Propensity Score Model Specification

The next step impacted by the decision to appraise treatment exposure at the subject or cluster level is determining the functional form of the PS model. When treatment exposure is appraised by subject, multiple approaches have been examined which are distinguished by their degree of complexity (Griswold et al., 2010; Leite et. al, 2015). The simplest approach is to fit a single-level PS model across the entire dataset, thereby completely ignoring any cluster dependency. That being said, cluster covariates, including aggregated covariates, can be included in the model as if they were subject

covariates. In applied research, this is the approach used by Belfi et al. (2016) and Wyse et al. (2008) when investigating the impact of school characteristics on student achievement. To account for cluster dependency, cluster fixed effects (i.e., dummy variables) can be added to the single-level model. Alternatively, a multilevel model with cluster random effects for intercept and/or slopes can be employed. This is the approach employed by Xiang and Tarasawa (2015). One critical distinction in their study was that students were clustered in elementary schools at baseline, but clustered in middle schools for treatment exposure (i.e., charter or public school) and the outcome measurement. For Belfi et al. (2016) and Wyse et al. (2008), baseline covariate and outcome measurement along with treatment exposure all occurred in the same set of schools. The simulation in the present study is more similar to Belfi et al. and Wyse et al., but the cluster structure complexity in Xiang and Tarasawa highlights yet another difficult methodological decision facing applied researchers with little guidance from current research.

When treatment exposure is appraised at the cluster level, the PS model specification decision is simplified to using a single-level model populated with cluster covariates. As noted, subject-level information can be incorporated in the estimation process via aggregated covariates. Most critically though, the decision to appraise treatment exposure by cluster results in all subjects within a cluster receiving the same PS. In contrast, if appraised by subject, subjects within the same cluster could potentially receive different PSs. This subsequently impacts the next step in a PS analysis.

Conditioning on the PS

The third step in a PS analysis is creating comparable treatment and control groups by conditioning on the PS. In reviews of PS studies with non-clustered data, the

most commonly used conditioning method is matching (Thoemmes & Kim, 2011; Zakrison et al., 2018). Matching methods have also been employed in a number of studies with clustered data where treatment exposure and outcome measurements were both at the subject level (e.g., Hughes, et al., 2010; Kim & Seltzer, 2007; McCormick et al., 2013). When treatment exposure is appraised at the subject level, a subject who received the treatment is matched with one or more subjects in the control condition who have a similar PS (Ho et al., 2007; Stuart, 2010). It is possible for treatment subjects to be matched with control subjects in the same or in different clusters. In contrast, when treatment exposure is appraised at the cluster level, the clusters are matched rather than the individual subjects. Simulation studies with clustered data and subject-level treatments have found matching methods to sufficiently minimize bias in the treatment effect estimation (Bellara, 2013; Leite, et al., 2015). There is, however, a critical drawback to matching methods. When an adequate match cannot be found for a subject (or cluster), they are dropped from the analysis. With treatment exposure by cluster, collecting data from a sufficient number of clusters can be difficult. The results from Chapter 2 showed that at least 60 clusters were needed in order to avoid estimation issues. Therefore, it is important to retain as many as possible, which potentially makes matching a suboptimal conditioning method.

An alternative conditioning method that largely alleviates the sample size issue is weighting on the PS (Austin & Stuart, 2015). In this method, the inverse of the PS is used as a weight in the subsequent estimation of the treatment effect. Given that a PS is estimated for every observation, the weight can also be calculated for every observation thereby retaining the original sample size. Furthermore, the process functions the same

way regardless of whether treatment is appraised at the subject or cluster level. The primary concern with weighting arises when the PS model is poorly specified resulting in erroneous PS estimates, and consequently, extremely large weights. Since weighting methods use the PS directly, they are more greatly impacted by PS model misspecification (Rubin, 2001). Contrastingly, matching methods are more robust to PS model misspecification given that the PS is used for the purpose of grouping treatment and control units rather than directly in the treatment effect estimation. That being said, extreme PS values might increase the difficulty of creating matched groups, thereby indirectly influencing treatment effect estimation.

In summary, matching and weighting methods each have their benefits and drawbacks. It is currently unclear how conditioning methods perform with a cluster-level treatment exposure and subject-level outcomes. Specifically, whether matching or weighting produces greater covariate balance between the treatment and control groups, thereby reducing selection bias, and minimizes bias in the treatment effect estimation. Furthermore, the ability of the two methods to balance aggregated covariates remains unexamined, although there is no a priori reason to expect any difference from non-aggregated covariates.

Present Study

When treatment exposure is a property of clusters, applied researchers conducting a PS analysis do not have the benefit of a random assignment mechanism to indicate whether treatment should be appraised at the cluster or subject level. Nonetheless, the decision has procedural consequences across multiple steps in the subsequent analysis. If treatment is appraised by subject, the PS is estimated for each subject with both subject

and cluster covariates likely needed in the model. Conditioning on the PS then also occurs by subject to create comparable treatment and control groups. In contrast, if treatment is appraised by cluster, subject covariates cannot be used directly to estimate the PS and must be aggregated to the cluster level. Estimation of and subsequent conditioning based on the PS then occurs by cluster with each subject within a cluster having the same PS. The present study employs a Monte Carlo simulation to explore the consequences of these procedural decisions on covariate balance and treatment effect estimation. The primary factor manipulated in the simulation is the appraisal of the treatment at the subject or cluster level in the PS analysis. The simulation also explores the utility of aggregated covariates as proxies for true cluster confounders missing from the analysis. Thus, when treatment exposure is a property of clusters at L2, but the outcome is measured by subjects at L1, the present study seeks to answer the following questions:

1. How is covariate balance impacted by the appraisal of a cluster-level treatment by subjects or clusters in the PS model and for conditioning on the PS?

2. When true cluster covariates are missing from the PS model and replaced by aggregated covariates, does covariate balance vary by the correlation between the aggregated and true covariates?

3. How is treatment effect estimation impacted by the appraisal of a cluster-level treatment by subjects or clusters in the PS model and for conditioning on the PS?

4. When true cluster covariates are missing from the PS and outcome models and replaced by aggregated covariates, does treatment effect estimation vary by the correlation between the aggregated and true covariates?

5. How do answers to the above questions differ by contextual factors, including the number of clusters, cluster size, cluster dependency of covariates to be aggregated and the outcome, and the use of matching or weighting as the PS conditioning method?

Method

Data Generation

Data were generated in *R* (v. 3.6.2; R Core Team, 2020) using the *simstudy* package (v. 0.1.15; Goldfeld, 2020). The R script to run the simulation is documented here: https://github.com/knickodem/AggCovsForPS. The procedures and desired sample characteristics in the present study align with those described in Chapter 2. To refresh, the samples produced in this simulation were two-level hierarchical structures with subjects, *i*, fully nested in clusters, *j*. This meant that each subject at L1 was nested within a single cluster at L2. Based on the results in Chapter 2, the Formative-RE procedure was chosen to generate the 30 covariates. The first step in this procedure was generating the 20 subject covariates (X_{1ij} - X_{20ij}) from a multivariate normal distribution with $\mu = 0$, variance σ^2 that varied by simulation condition, and correlated at $\rho = .20$. Next, a cluster adjustment drawn from a distribution of $N(0, \tau_{00})$ was added to the subject covariates to induce cluster dependency. The mean for each of the first 10 subject covariates ($X_{1ij} - X_{10ij}$) was calculated within each cluster *j* to create the 10 true cluster covariates ($\bar{X}_1$ - $\bar{X}_{10}$). These values were used to generate the true probability of treatment exposure (i.e., true PS), and the subject-level outcome Y_{ij}. With σ^2 varying by simulation condition, error drawn from $N(0, \sigma^2)$ was added to the 10 true cluster covariates ($\bar{X}_{1j} - \bar{X}_{10j}$) for each cluster to create the 10 aggregated covariates ($\bar{X}'_{1j} - \bar{X}'_{10j}$).

Treatment exposure, Z_j, was created as a property of clusters. The 10 true L2 covariates were used to calculate the true PS for each cluster, *j*, with a single-level model. This is the same as Equation 2 in Chapter 2:

$$\text{logit}(Z_j = 1) = \beta_0 + \beta_1\bar{X}_{1j} + \beta_2\bar{X}_{2j} + \beta_3\bar{X}_{3j} + \beta_4\bar{X}_{4j} + \beta_5\bar{X}_{5j} + \beta_6\bar{X}_{6j} + \beta_7\bar{X}_{7j} + \beta_8\bar{X}_{8j} + \beta_9\bar{X}_{9j} + \beta_{10}\bar{X}_{10j} + u_j \quad (2)$$

The regression coefficients β_1 - β_{10} were fixed to 0.50. The intercept, β_0, was set to -1.386294 to impose a marginal probability of treatment exposure (i.e., mean PS) to .20. Based on the results of Chapter 2, this was intended to produce a treatment-to-control ratio of approximately 1:3. The cluster residuals, u_j, were drawn from a logit distribution with $\mu = 0$ and $\sigma^2 = \pi^2/3$. Doing so produced overlapping probability of treatment distributions of the treatment ($Z_j = 1$) and control groups ($Z_j = 0$). The two groups were determined with Equation 3 from Chapter 2:

$$(Z_j = 1) \textit{ if } \text{logit}(Z_j = 1) > 0, \textit{ else } (Z_j = 0) \quad (3)$$

The true outcome, Y_{ij}, for each subject *i* in cluster *j* was generated with the same two-level random intercept model used in Chapter 2 (Equation 4):

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X_{5ij} + \beta_{6j}X_{6ij} + \beta_{7j}X_{7ij} + \beta_{8j}X_{8ij} + \beta_{9j}X_{9ij} + \beta_{10j}X_{10ij} + \beta_{11j}X_{11ij} + \beta_{12j}X_{12ij} + \beta_{13j}X_{13ij} + \beta_{14j}X_{14ij} + \beta_{15j}X_{15ij} + \beta_{16j}X_{16ij} + \beta_{17j}X_{17ij} + \beta_{18j}X_{18ij} + \beta_{19j}X_{19ij} + \beta_{20j}X_{20ij} + r_{ij} \quad (4)$$

$$r_{ij} \sim N(0, \sigma^2)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{X}_{1j} + \gamma_{02}\bar{X}_{2j} + \gamma_{03}\bar{X}_{3j} + \gamma_{04}\bar{X}_{4j} + \gamma_{05}\bar{X}_{5j} + \gamma_{06}\bar{X}_{6j} + \gamma_{07}\bar{X}_{7j} + \gamma_{08}\bar{X}_{8j} + \gamma_{09}\bar{X}_{9j} + \gamma_{0\,10}\bar{X}_{10j} + \gamma_{0\,11}Z_j + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

…

$\beta_{20j} = \gamma_{20\,0}$

The subject level, r_{ij}, and cluster level, u_{0j}, error terms were drawn from a normal distribution with $\mu = 0$. The variances of σ^2 and τ_{00}, respectively, varied depending on the simulation condition. Other than the intercept, γ_{00}, which was set to 0, the regression coefficients were fixed to 0.50. This included the treatment effect, $\gamma_{0\,11}$, implying that the effect was homogenous across all PSs. In this study, the 10 cluster covariates were true confounders as they were related to both treatment exposure (Equation 2) and the outcome (Equation 4). In contrast, the 20 subject covariates were only associated with the outcome (Equation 4).

In light of the results from Chapter 2, the Formative-RE procedure produced sample datasets where the average correlation between the true and the aggregated cluster covariates ranged from approximately .20 to .98 depending on the magnitude of error added to the true values. The cluster dependency (ICC(1)) and reliability (ICC(2)) of the aggregated values varied as expected given the manipulations of ICC(1) and subjects per cluster. Likewise, the cluster dependency of the outcome *Y* varied by ICC(1) condition. Lastly, the correlation among the subject covariates remained close to the imposed correlation of .20 while the correlation among the true cluster covariates and aggregated cluster covariates was < .05.

Manipulated Factors

Six simulation factors were manipulated: number of clusters, number of subjects per cluster, ICC(1) of the aggregated covariates and the outcome variable, magnitude of error in the aggregated cluster covariates, treatment appraisal at the subject or cluster

level, and PS conditioning method (Table 3.1). Employing a 3 x 3 x 3 x 4 x 2 x 2 fully crossed factorial design, the study had a total of 432 conditions. Each condition was replicated 1000 times.

Table 3.1

Summary of Simulation 2 Manipulated Factors and Levels

Factor	Levels
Number of clusters	60, 100, 140
Number of subjects per cluster	20, 60, 100
ICC(1) of outcome & aggregated covariates	.05, .10, .20
Aggregated L2 error magnitude	σ of random error: 0.1, 0.3, 0.5, 1.0
Treatment appraisal	Cluster Level, Subject Level
Conditioning method	Matching, Weighting

Contextual Factors. The first three factors were chosen because they are often of concern for applied researchers when determining how to conduct a PS analysis. The Chapter 2 results revealed the simulation design used here required a minimum of 60 clusters for the PS estimation model to converge with consistency. In the field of education, clusters are often classrooms, schools, and districts. Gathering data on a large number of clusters can be challenging and expensive. To strike a balance between practical necessity and realistic scenarios, the present simulation examined conditions with 60, 100, and 140 clusters. The number of subjects per cluster [20, 60, 100] and the ICC(1) for the aggregated covariates and the subject-level outcome [.05, .10, .20] remained the same as in Chapter 2. These represent values commonly found in education research. The fourth factor was the magnitude of error in the aggregated covariates. Manipulating the error magnitude aids in determining the point at which an aggregated covariate becomes an unreliable indicator for the true cluster covariate in a PS analysis. The levels examined were the same as in Chapter 2. Specifically, the standard deviation

of the error added a true cluster values to generate the aggregated covariate values was 0.1, 0.3, 0.5, or 1.0.

Treatment Appraisal. Treatment exposure was generated as a cluster-level property, but could be appraised by cluster (L2) or by subject (L1) in the specification of the PS estimation model and when conditioning on the PS. In the cluster appraisal condition the PS model was the same as Equation 2, but used the aggregated cluster covariates ($\bar{X}'_1$ - $\bar{X}'_{10}$) instead of the true cluster covariates ($\bar{X}_1$ - $\bar{X}_{10}$). The PS was estimated by cluster rather than using the full dataset with each subject-level observation. Consequently, every subject within a cluster had the same PS, so conditioning on the PS also occurred by cluster. In the subject appraisal condition, the PS was estimated for each subject using a single-level model represented by:

$$\text{logit}(Z_{ij} = 1) = \beta_0 + \sum_{p=1}^{20} \beta_p X_{pij} + \sum_{q=1}^{10} \gamma_q \overline{X}'_{qj} \quad (9)$$

where X_{pij} was a vector of the 20 subject covariates, p, for subject i in cluster j. The 10 aggregated covariates, $\bar{X}'_{qj}$, were also included, but regarded as subject characteristics in the model. In both the cluster and subject appraisal conditions, the outcome model was the same as Equation 4, but once again used the aggregated covariates ($\bar{X}'_1$ - $\bar{X}'_{10}$) as proxies for the true cluster covariates ($\bar{X}_1$ - $\bar{X}_{10}$). Conceptually, the outcome model in the subject appraisal condition would include treatment at L1 ($\beta_{21j} Z_{ij}$) instead of at L2 ($\gamma_{011} Z_j$), but for practical purposes the outcome models in the two conditions were equivalent. The single-level PS models were estimated using the glm function in base *R* while the multilevel outcome model was estimated using the *lme4* package (v. 1.1.21; Bates et al., 2015).

Initially, two other PS models were proposed that also conceptualized treatment exposure as a subject-level property. The first was a cluster fixed effects model which was a single-level with *J*-1 dummy variables to account for potential cluster dependency:

$$\text{logit}(Z_{ij} = 1) = \beta_0 + \sum_{p=1}^{20} \beta_p X_{pij} + \sum_{j=1}^{J-1} \delta_j G_j \quad (10)$$

where *G* was the vector of *J*-1 dummy variables. The other model was a cluster random effects model represented by:

$$\text{Level 1: logit}(Z_{ij} = 1) = \beta_{0j} + \sum_{p=1}^{20} \beta_{pj} X_{pij} \quad (11)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \sum_{q=1}^{10} \gamma_{0q} \overline{X}'_{qj} + u_{0j} \qquad u_{0j} \sim N(0, \tau_{00})$$

…

$$\beta_{pj} = \gamma_{p0}$$

where u_{0j} was the random intercept for cluster *j*. In PS models, the criterion is treatment exposure. With treatment exposure at the cluster level, the dummy variables for clusters in Equation 10 or the random effects for clusters in Equation 11 become perfect predictors of the criterion. Thus, any PS model including an indicator for cluster will fail to converge because the PS can be determined to be either 1 or 0 using the cluster indicator. In contrast, the models in Equations 2 and 9 only included cluster-level covariates, but not an indicator for cluster. Consequently, the PS from these models can be calculated probabilistically rather than deterministically, which satisfies the overlap assumption of PS methods. For Equations 10 and 11 to be viable in a PS model with a cluster-level treatment, the clusters where treatment exposure occurs must be different than the clusters represented by the fixed and random effects.

Conditioning Method. The treatment and control groups were equated using either a matching or a weighting method. Matching and weighting were performed on clusters in the cluster appraisal condition and by subjects in the subject appraisal condition. To maintain the overlap assumption, following estimation of the PS, observations with a PS > .999 or PS < .001 were removed. In the weighting condition, this had a side effect of limiting extreme weights by imposing a maximum weight of 1000. The weights were calculated via the *WeightIt* package (v. 0.6.0; Greifer, 2019) from the inverse of the estimated PS ($\widehat{PS}$) using the formula:

$$w_i = \frac{Z_i}{\widehat{PS}_i} + \frac{(1 - Z_i)}{1 - \widehat{PS}_i} \quad (12)$$

where *Z* is the treatment indicator and *i* represented subjects in the subject appraisal condition and clusters in the cluster appraisal condition. Equation 12 calculates the weights for estimating the average treatment effect. Although the comparison matching method could only estimate the average treatment effect on the treated, the two estimates were expected to be identical given that the true effect was constant across all subjects in the current simulation design (Ho et al., 2007). The weights were then used as precision weights in the outcome model run in *lme4* (Bates et al., 2015), which does not accept sampling weights. Treating the weights from the PS model as precision or sampling weights should not impact estimation of the treatment effect, but could differentially impact standard errors (Snijders & Bosker, 2012).

In the matching condition, treatment and control groups were matched on the logit of the PS with the *MatchIt* package (v. 3.0.2; Ho et al., 2011). Nearest neighbor greedy matching was utilized on a 1:1 basis without replacement and a caliper width of 0.2 standard deviations of the logit. Although many matching strategies are available, these

settings are amongst the most commonly used (Thoemmes & Kim, 2011; Zakrison et al., 2018). In the subject appraisal condition, treatment subjects could be matched with control subjects within the same cluster or from a different cluster.

Dependent Variables

To answer the research questions, the first concern is whether the estimation models even converge. All of the PS and outcome models utilize maximum likelihood estimation which relies on an iterative procedure to estimate the model parameters that maximize the likelihood function (Raudenbush & Bryk, 2002). When the maximum cannot be found within the specified number of iterations, model estimation is non-convergent. A myriad of possible factors can contribute to convergence issues, including model misspecification or inadequate sample size. Ultimately, non-convergence is a sign of untrustworthy parameter estimates. Non-convergence rates were tracked across conditions to identify problematic specifications of the PS and outcome models.

Following estimation of the PS, one of the previously described conditioning methods was used to create comparable treatment and control groups. One disadvantage of matching methods is some subjects in the treatment group may not have a suitable match in the control group. Unmatched units are then dropped from the subsequent analysis. In contrast, weighting methods often utilize the entire sample when estimating the treatment effect in the outcome model. Therefore, I tracked the number of matched subjects and clusters to aid in comparing the two PS conditioning methods. This also gives an indication of whether the PS model produced estimates close to 0 or 1, and thus, were removed to maintain the overlap assumption.

From the results in Chapter 2 we know the characteristics of the sample datasets initially generated. However, after conditioning on the PS, the samples actually used to estimate the treatment effect in the outcome model may have different characteristics. Of primary interest were the characteristics of the 10 aggregated covariates. To put the treatment effect estimation in the proper context, I calculated the correlation between the 10 true L2 covariates and the aggregated counterparts[4], the reliability (ICC(2)) of the aggregated covariates at the cluster level, and cluster dependency (ICC(1)) of the aggregated covariates at the subject level for the PS samples.

To determine the comparability of the treatment and control groups, covariate balance between the two groups on the 20 subject covariates and 10 cluster covariates was checked using the absolute standardized difference (Austin, 2011). For each covariate this was calculated as the absolute value of the difference in sample means, $\bar{X}$, between the treatment and control groups divided by a standard deviation (SD):

$$\text{ASD} = | (\bar{X}_{\text{treatment}} - \bar{X}_{\text{control}}) / \text{SD} | \quad (13)$$

Specifically, the pooled SD from the treatment and control group in the full sample prior to conditioning was used here (Stuart, 2010). The absolute standardized difference was chosen over other balance evaluation techniques (e.g., c-statistic, overlapping coefficient, t-test) because simulation studies have shown it to more often select the correctly specified PS model (Ali, et al., 2014; Austin et al., 2007). Additionally, the absolute standardized difference is not influenced by sample size and can be used equivalently for subject- and cluster-level covariates.

[4] The average correlation across the 10 pairs of true and aggregated L2 covariates was calculated by first transforming the correlations to z-scores using Fisher's r-to-z transformation (Silver & Dunlap, 1987). The z-scores were transformed back to correlations after averaging across the 1000 replications.

When the absolute standardized difference was < 0.10, the covariate was considered balanced. Thresholds as low as 0.05 (WWC, 2020) and as high as 0.25 (Rubin, 2001) have been suggested. Nonetheless, 0.10 is generally considered a suitable threshold for both continuous and binary covariates (Austin, 2009; Kainz et al., 2017). Using the 0.10 threshold, the count of covariates balanced was used to evaluate the impact of subject level or cluster level treatment appraisal and use of aggregate covariates. The ratio of a covariate's variance in the treatment and control group is another indicator of balance for continuous covariates. The two groups are balanced if the ratio is close to 1.0. Therefore, the count of covariates with a variance ratio < 2.0 was also tracked (Rubin, 2001). Additionally, across all covariates the mean absolute standardized difference and mean variance ratio were calculated to provide a summary of the balance for each simulation condition.

The treatment effect estimation was evaluated by calculating bias, mean absolute error (MAE), and root mean square error (RMSE). With the true treatment effect set to 0.50, bias was calculated across the 1000 replications by:

$$\text{Bias} = \frac{\sum_{n=1}^{1000}(\hat{\delta} - 0.50)}{1000} \tag{14}$$

where $\hat{\delta}$ is the estimated average treatment effect. Positive bias indicates systematic overestimation of the treatment effect while negative bias suggests systematic underestimation. MAE and RMSE are indicators of precision in the estimation. MAE weights each observation equally when calculating the overall magnitude of error while RMSE gives greater weight to large errors. MAE and RMSE were calculated across the 1000 replications.

$$\text{MAE} = \frac{\sum_{n=1}^{1000}|\hat{\delta} - 0.50|}{1000} \quad (15)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{1000}(\hat{\delta} - 0.50)^2}{1000}} \quad (16)$$

Lastly, a set of baseline measures were calculated from the full sample to which the efficacy of the PS methods under consideration could be judged. Utilizing the full sample without PS methods was akin to conducting an observational study with a multilevel outcome model to estimate the treatment effect while adjusting for the subject and cluster covariates, as well as adjusting the standard errors for cluster dependency. To establish baseline measures of covariate balance, the mean absolute standardized difference, mean variance ratio, and count of covariates with acceptable absolute standardized difference and variance ratio were calculated. If the outcome model converged using a PS method, the treatment effect was then estimated using the full sample. The baseline convergence rate, bias, MAE, and RMSE were then recorded.

Analytic Procedure

With the exception of bias, MAE, and RMSE, which were calculated using Equations 14 – 16, the mean for each dependent variable was calculated across the 1000 replications for each of the 432 conditions. As in Chapter 2, partial-ω^2 was calculated from a factorial ANOVA on each dependent variable to determine the amount of variance explained by each of the six independent and all two-way interactions. Model assumptions were checked for each factorial ANOVA using a Q-Q plot of the residuals, scatterplot of residuals and predicted values, and calculation of variance inflation factor for each predictor (Williams et al., 2013). Lastly, boxplots were created for visual inspection of the variation in the dependent variables by select manipulated factors.

Results

As in Chapter 2, the Q-Q plot and scatterplot revealed non-normality and /or heteroscedasticity in the residuals of the factorial ANOVA model for the majority of the dependent variables. Consequently, the partial- ω^2 values were treated as a crude indicator of the association between a dependent variable and the manipulated factors (Appendix B). Strict interpretation of the values as the variance of a dependent variable explained by a manipulated factor relative to the other factors is not recommended. The conclusions for the study instead relied on descriptive comparisons.

Convergence

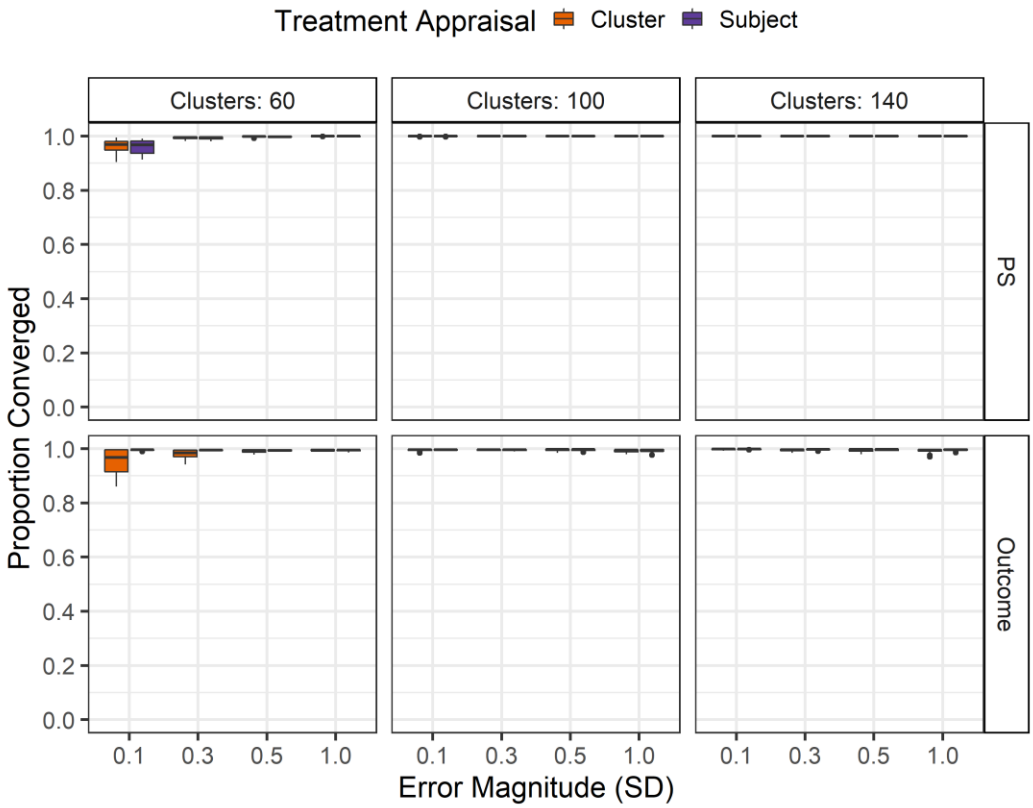
Convergence of the PS and outcome models is a necessary antecedent for producing and interpreting the parameter estimates from the models. For both the PS and outcome models, there were little to no convergence issues in the 100 and 140 cluster conditions (Figure 3.1). With 60 clusters issues arose in the smallest error magnitude condition. The median convergence rate for the PS model was 96.8%, though the minimum rate was 90.5%. The rate improved as the error magnitude in the aggregated covariates increased. The additional error made the PS estimation less deterministic, and thus, more likely to converge. This trend, however, has little bearing on practice unless the PS estimation model is identical to the true PS model, which is unlikely.

The outcome model was estimated only for PS models that converged. Yet, some convergence issues lingered for the 60 cluster condition, particularly with treatment appraised at the cluster level. In this condition, the median convergence rate was 99.2%, but the minimum was only 86.1%. In contrast, the minimum was 98.6% with subject appraisal and 60 clusters. For comparison, the baseline model - which estimated the

treatment effect using the full dataset - had a minimum convergence rate of 96.9% in the 60 cluster condition. The convergence rates in the PS and outcome models once again highlights the need for a sufficient number of clusters regardless of the number of subjects per cluster and the appraisal of treatment at the cluster or subject level.

Figure 3.1

Convergence Rate for Propensity Score (PS) and Outcome Models



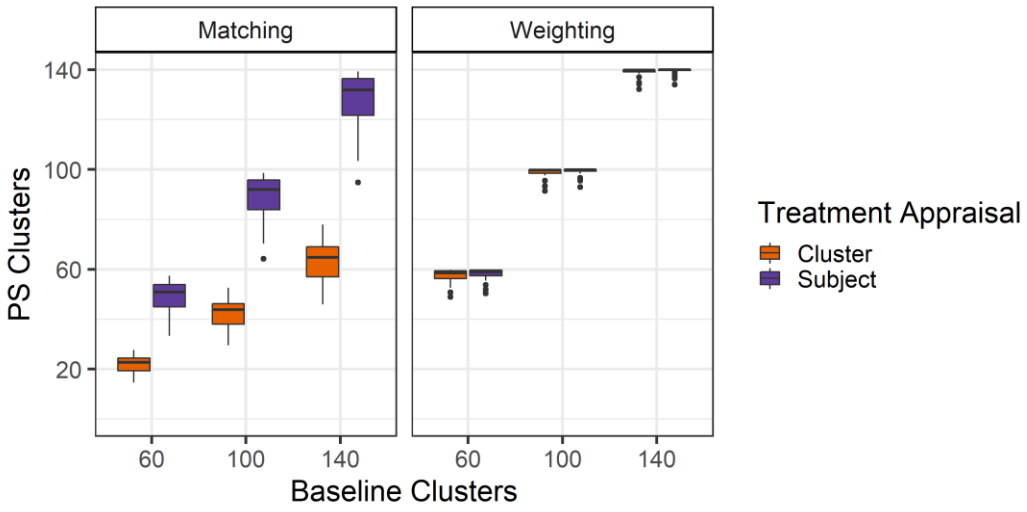
Sample Characteristics After Conditioning on the PS

Sample Size. The proportion of treated subjects in the baseline datasets ranged as expected from .26 to .34. This is equivalent to treatment-to-control ratios between approximately 1:3 and 1:2. The matching or weighting procedure then created the treatment and control groups used for estimating the treatment effect. As expected, the

weighting condition retained a similar number of clusters as the full baseline sample (Figure 3.2). This indicated that for the PS models that converged, few extreme weights were removed. In the matching condition, treatment appraisal by subjects retained slightly fewer clusters than its counterpart in the weighting condition. Treatment appraisal by clusters, however, retained far fewer clusters in the analytic sample. In the 140 cluster condition, subject appraisal retained a median of 132 clusters whereas cluster appraisal only retained 65. In the 60 cluster condition, the cluster appraisal retained as few as 14 clusters, meaning 7 treatment clusters were matched to 7 control clusters.

Figure 3.2

Number of Clusters Retained From Conditioning on the Propensity Score (PS)

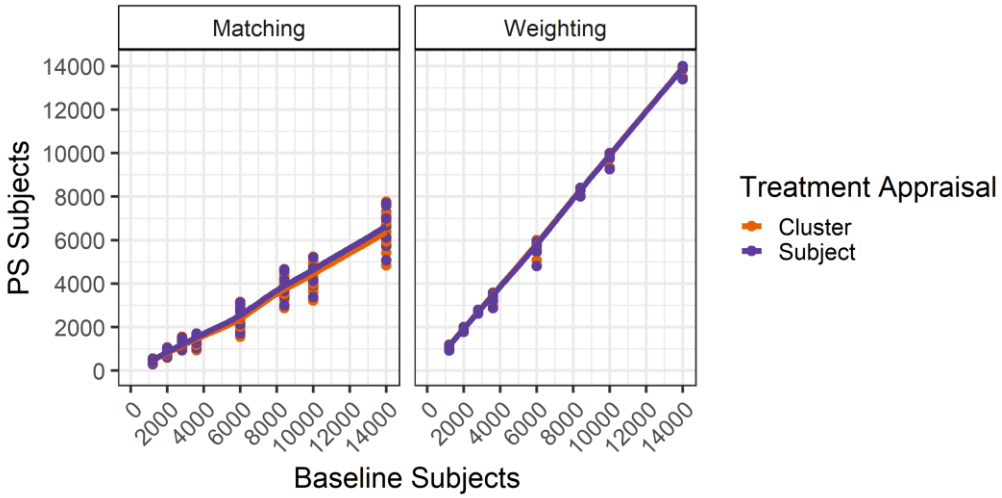


Despite the difference in matched clusters by treatment appraisal, there was little difference in the number of subjects retained (Figure 3.3). When appraised by cluster, all subjects within a cluster were retained when the cluster was successfully matched. When appraised by subject, however, the median sampling ratio was .50. This meant half the subjects within a cluster were successfully matched while the other half were not.

Compared to the weighting condition, the matching condition produced substantially smaller samples regardless of treatment appraisal level. In the smallest condition with 60 clusters and 20 subjects per cluster (n = 1200), the median number of subjects in the weighting condition was 1154.3, but only 460.7 for matching.

Figure 3.3

Number of Subjects Retained From Conditioning on the Propensity Score (PS)



Aggregated Covariate Characteristics. To make appropriate comparisons it is important to know whether the reduced samples in the matching condition were still representative of the full baseline samples. For the present study, the characteristics of the aggregated covariates as measured by ICC(1), ICC(2), and the correlation between the aggregated and true cluster covariates were of particular interest. ICC(1) - the measure of cluster dependency in the 10 aggregated covariates – was manipulated to be .05, .10, and .20. In the matched samples the median ICC(1) was .05, .09, and .19, respectively. The weighted samples produced values of .05, .10, and .20, respectively. Likewise, the matched and weighted samples yielded similar values for ICC(2) – reliability of the

aggregated covariates – which varied as expected given the imposed ICC(1) and subjects per cluster conditions (Table 3.2).

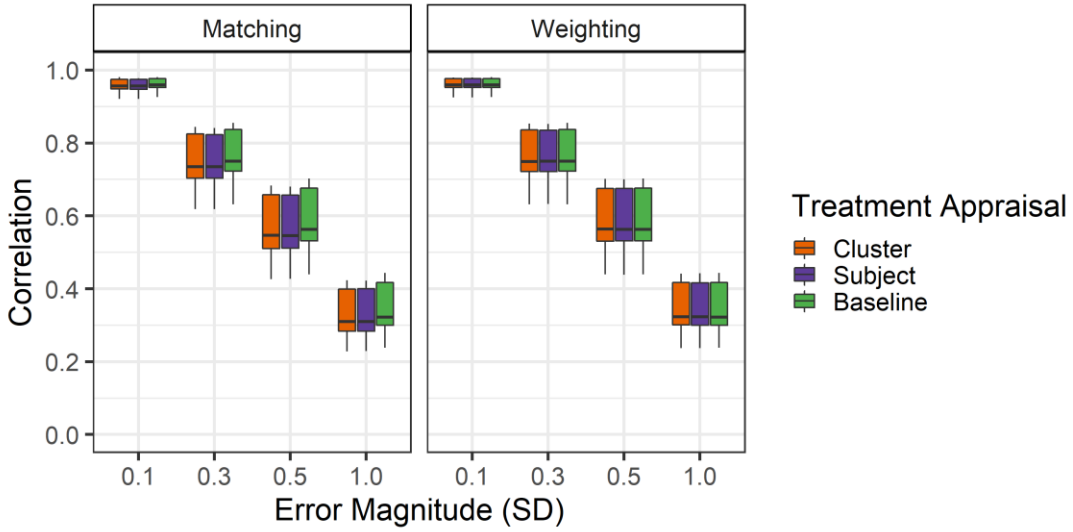
Table 3.2

Expected and Propensity Score Sample ICC(2) Values

Condition		Expected	Matching		Weighting	
ICC(1)	# Subjects	ICC(2)	Median	Range	Median	Range
.05	20	.51	.45	.39 - .50	.50	.47 - .51
.05	60	.76	.74	.70 - .75	.75	.75 - .76
.05	100	.84	.83	.80 - .84	.84	.83 - .84
.10	20	.69	.64	.58 - .68	.68	.66 - .69
.10	60	.87	.85	.83 - .87	.87	.86 - .87
.10	100	.92	.91	.89 - .91	.92	.91 - .92
.20	20	.83	.80	.76 - .83	.83	.81 - .83
.20	60	.94	.93	.91 - .94	.94	.93 - .94
.20	100	.96	.96	.94 - .96	.96	.96 - .96

Figure 3.4

Mean Correlation Between 10 True and Aggregated Cluster Covariates



Lastly, the correlations between the aggregated and true cluster covariates in the PS samples were similar to the correlations in the full baseline sample regardless of the conditioning method or level of treatment appraisal (Figure 3.4). As the imposed standard

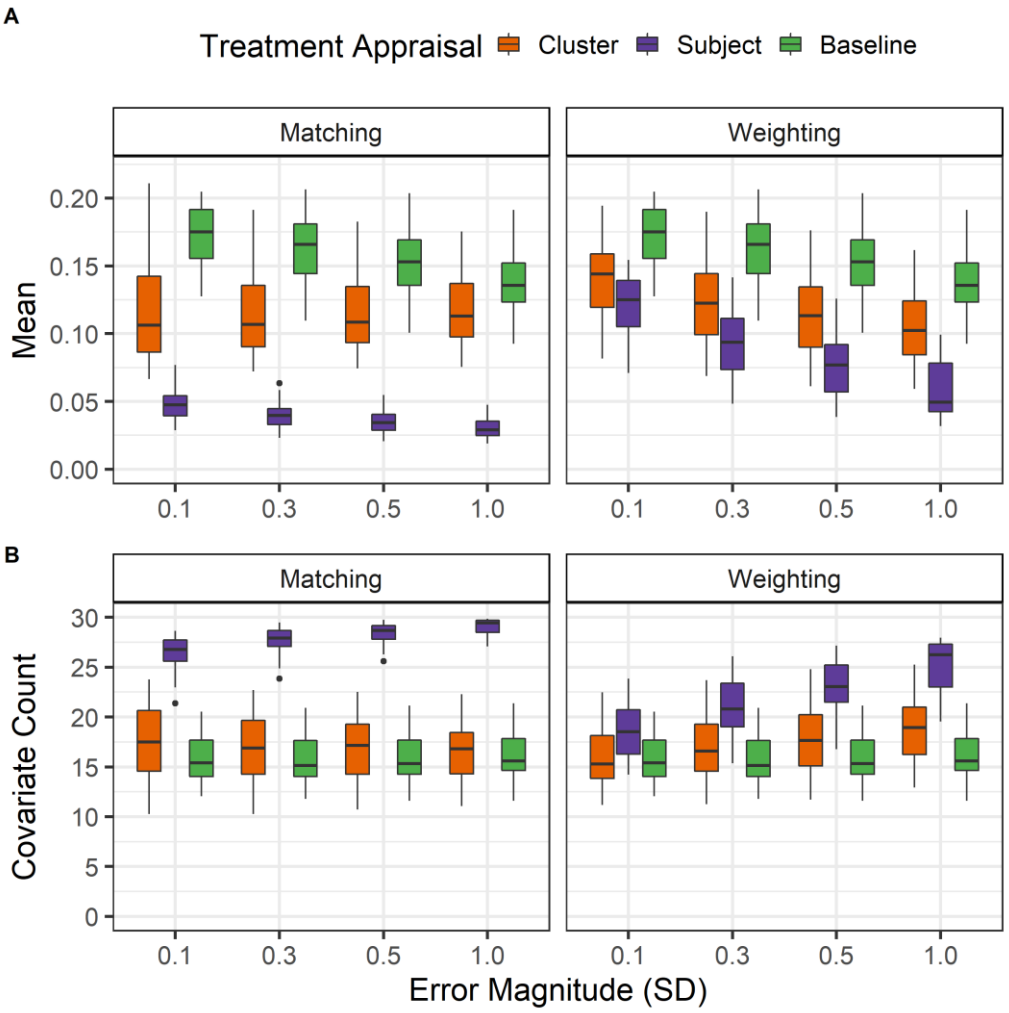
deviation of the error magnitude increased between 0.1, 0.3, 0.5, and 1.0, the range of the correlations decreased between .92 - .98, .62 - .85, .43 - .70, and .23 - .44. These results aligned with those from Chapter 2. The systematic variation in correlation by error magnitude enables the evaluation of aggregated covariates as replacements for the true cluster covariates. Furthermore, the similarity of the correlation, ICC(1), and ICC(2) across treatment appraisal and PS conditioning conditions allows for appropriate comparison of aggregated covariate utility between these conditions.

Covariate Balance

Absolute Standardized Difference. The first research question explored whether covariate balance was impacted by appraisal of treatment by cluster or subject in the specification of the PS model and subsequent conditioning method. When treatment and control groups are balanced on the covariates, selection bias due to those covariates is neutralized. Results show there was greater covariate balance as measured by the mean absolute standardized difference with subject level appraisal than cluster level appraisal (Figure 3.5A). When appraised by cluster, the mean absolute standardized difference was similar for both matching (*Mdn* = .11) and weighting (*Mdn* = .12), and generally larger than the accepted standard of 0.10. When appraised by subject, matching (*Mdn* = .04) consistently outperformed weighting (*Mdn* = .08). That being said, weighting still produced acceptable covariate balance except in the smallest error magnitude condition. Overall, the PS methods demonstrated greater covariate balance, thereby a reduction in selection bias, compared to the baseline covariate balance from the full unadjusted sample.

Figure 3.5

Mean Absolute Standardized Difference (A) and Count of Covariates with a Difference <0.10 (B) for 20 Subject and 10 Cluster Covariates



Beyond the overall mean absolute standardized difference, the treatment and control groups were considered balanced if the absolute standardized difference was < 0.10 for each of the 30 covariates (Figure 3.5B). The median count of balanced covariates for the unadjusted baseline measure was 15.3. Covariate balance did not improve much over baseline when treatment was appraised by cluster regardless of whether the matching (*Mdn* = 17.2) or weighting (*Mdn* = 17.2) method was used. The covariate

balance increased with subject-level appraisal, but was still largely unsatisfactory in the weighting condition (*Mdn* = 22.2). Only the subject-level treatment appraisal paired with the matching method (*Mdn* = 28.2) produced covariate balance suggesting selection bias in the 30 covariates was largely removed from the treatment effect estimation.

The second research question explored if covariate balance differed by the quality of the aggregated covariates. Generally, as the error magnitude increased, covariate balance improved in the PS samples and in the full baseline samples (Figure 3.5A). For instance, in the subject appraisal condition with the weighting procedure, the mean absolute standardized difference between the treatment and control groups was above the accepted threshold of 0.10 at the smallest error magnitude (*Mdn* = 0.13), but well under the threshold at the largest error magnitude (*Mdn* = 0.05). Based on the data generation procedure, the treatment group had systematically higher values for the true cluster covariates than the control group. This pattern implies that the added error in the aggregated covariates increased the overlap in the distributions between the treatment and control groups. In other words, balance between the treatment and control groups improved as the aggregated covariates became less accurate proxies for the true covariates. The pattern of covariate balance improving as error magnitude increased remained for the count of balanced covariates in the subject appraisal condition, but was less pronounced in the cluster appraisal condition and non-existent in the baseline samples (Figure 3.5B).

Regarding contextual factors, the mean absolute standardized difference tended to improve as the number of clusters rose from 60 (*Mdn* = .13) to 100 (*Mdn* = .09) to 140

(*Mdn* = .08). Likewise, the count of balanced covariates also improved as the number of clusters increased from 60 (*Mdn* = 17.2) to 100 (*Mdn* = 20.4) to 140 (*Mdn* = 22.6).

Variance Ratio. Covariate balance can also be evaluated by comparing the ratio of the variances in continuous covariates between the treatment and control groups. Variance ratios < 2.0 were considered acceptable. Across all conditions the variance ratio did not exceed 1.5 with the exception of matching on the PS from cluster-level treatment appraisal with 60 clusters. Although the median of 1.4 was still acceptable, 12 of the 36 observations from this combination of factors had values > 2.0. Regarding the 30 covariates, the count of balanced covariates was high with the baseline samples (*Mdn* = 29.6), treatment appraised by cluster (*Mdn* = 28.8), and treatment appraised by subject (*Mdn* = 29.0).

Regarding the second research question, there was a small but consistent trend that as the error magnitude increased, the variance ratio improved slightly both overall and by count of covariates. At the smallest error magnitude the median variance ratio was 1.25 (*Mdn* count = 28.3); at the largest error magnitude the median improved to 1.18 (*Mdn* count = 29.3).

Taken as a whole, the covariate balance results provide some insights for applied researchers. Appraising treatment at the subject-level paired with the matching method produced satisfactory covariate balance across all indicators. In other words, selection bias was adequately reduced. Conversely, imbalance between the treatment and control groups remained in the other conditions. Regarding aggregated covariates, greater balance was achieved with lower correlations between the aggregated covariate and its true counterpart. Thus, the desire for a closely associated replacement for a missing true

cluster confounder may come at the expense of greater difficulty producing equivalent treatment and control groups. Lastly, PS samples with more clusters tended to have better balance.

Treatment Effect Estimation

The simulation results revealed few differences in treatment effect estimation between the PS methods. Rather, the quality of aggregated covariates as replacements for the missing true cluster confounders had a greater impact on estimation of the treatment effect. At the smallest error magnitude – where the correlation between aggregated and true cluster covariates was .92 - .98 – the median bias was 0.04 (Figure 3.6A). This constitutes a slight overestimation of the true treatment effect, which was set to 0.50. Furthermore, with the sole exception of matching with treatment appraisal by cluster and 60 clusters (SD = 0.13), there was little variation in the bias (SD < 0.01). When the correlation dropped to .62 – .85, still a reasonably high range of correlations, the median bias increased to 0.23 (SD = 0.04) across the PS conditions. A bias of 0.23 indicates the PS methods, on average, estimated the treatment effect to be 46% larger than the true effect. Bias continued to increase as the correlation between the aggregated and true covariates decreased.

A similar pattern emerged for the estimation precision as measured by MAE (Figure 3.6B). At the smallest error magnitude, in the matching with 60 cluster and cluster treatment appraisal condition the median MAE (Mdn = 0.55) was higher and variation in MAE (SD = 0.21) larger than other PS methods (Mdn = 0.10, SD = 0.03). Thus, when the aggregated cluster covariates were correlated with the missing true covariates in a range of .92 - .98, the MAE of 0.10 signifies that the mean difference

between the true treatment effect of 0.50 and estimated effect was 0.10. As the correlations decreased to a range of .62 - .85, the estimation precision also decreased as indicated by the higher MAE values (*Mdn* = 0.24, *SD* = 0.10). RMSE, which gives greater weight to large errors, showed the same pattern with more emphasis on the flaws in matching with 60 clusters and treatment appraisal by cluster (Figure 3.6C).

Figure 3.6

Treatment Effect Estimation Bias(A), Mean Absolute Error (B), and Root Mean Square Error (C) by Propensity Score Method and Error in the Aggregated Covariates

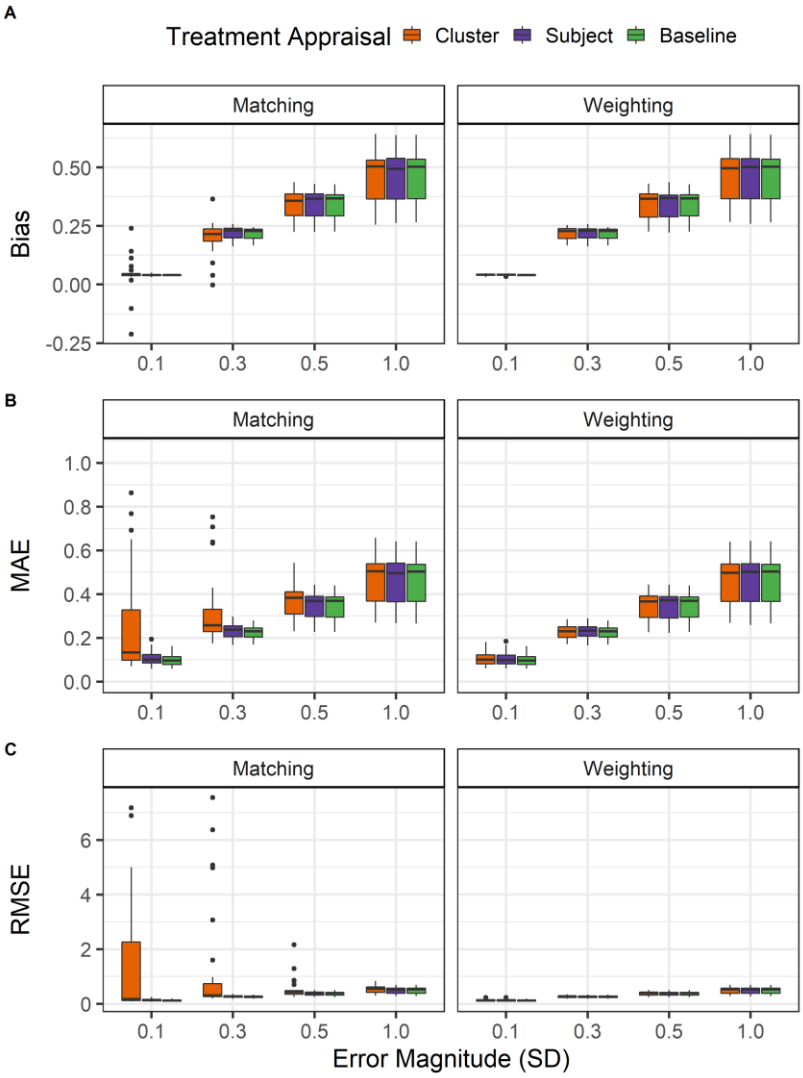
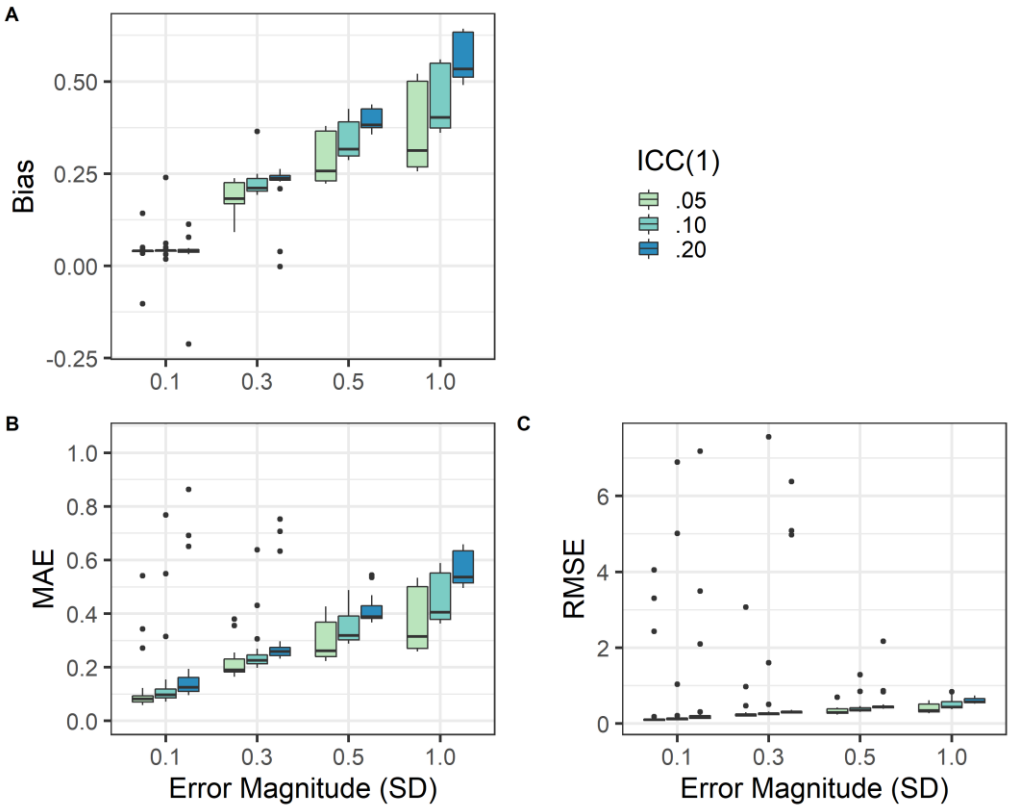


Figure 3.7

Treatment Effect Estimation Bias(A), Mean Absolute Error (B), and Root Mean Square Error (C) with Propensity Score Methods by Cluster Dependency (ICC(1)) Condition



In addition to differences in estimation bias and precision with 60 clusters compared to 100 or 140 clusters, the other contextual factor impacting treatment effect estimation was the cluster dependency of the aggregated covariates and the subject-level outcome as measured by ICC(1). Overall, bias and error increased as ICC(1) increased for all PS methods (Figure 3.7). Furthermore, there was an interaction with error magnitude in the aggregated covariates. At the smallest error magnitude, the largest cluster dependency condition of ICC(1) = .20 had a similar bias (*Mdn* = 0.04, *SD* = 0.05) to the smallest when ICC(1) = .05 (*Mdn* = 0.04, *SD* = 0.03). However, at the largest error

magnitude, bias for ICC(1) = .20 (*Mdn* = 0.54, *SD* = 0.06) was substantially higher than for the ICC(1) = .05 condition (*Mdn* = 0.31, *SD* = 0.10). The same pattern held for the measures of MAE and RMSE. Thus, poor estimation of the treatment effect when aggregated covariates were inferior proxies for missing true cluster covariates was compounded in the presence of higher cluster dependency. This trend is particularly interesting given that higher cluster dependency of the aggregated covariate values leads to higher reliability for the aggregated covariates (see Table 3.2).

The treatment effect estimation results revealed possible guidance for applied researchers when investigating the impact of a cluster-level treatment exposure on subject-level outcomes. With fewer than 100 clusters, appraising treatment at the cluster level in the PS model in conjunction with conditioning via matching produces inferior estimation accuracy to other PS methods. With more than 100 clusters the choice of treatment appraisal level and conditioning method is far less consequential than the quality of the aggregated covariates used in the analysis. Only aggregated covariates highly correlated (*r* = .92 - .98) with the missing true cluster covariates yielded acceptable estimation. The inaccurate estimation with weakly correlated aggregated covariates was exacerbated as cluster dependency in the aggregated covariates and subject-level outcome increased. A possible implication is that the efficacy of an aggregated covariate as a replacement for a missing true cluster confounder depends more on the high correlation between the true and aggregated covariate than the reliability of the aggregated covariate. That being said, the simulation design more closely imitated a formative aggregation, so the reliability of the aggregated values may not be meaningful in this context.

Discussion

When treatment exposure occurs by cluster, yet subject-level outcomes are of interest, the literature provides little guidance for navigating the numerous decisions for conducting a PS analysis. The simulation in the present study sought to provide applied researchers with evidence for appraising treatment at the cluster or subject level when specifying the PS estimation model and conditioning on the PS. The primary conclusions drawn from the simulation results are: 1) Covariate balance and treatment effect estimation improved with more clusters. More than 60 clusters, and ideally at least 100 clusters, are needed to minimize model non-convergence. 2) Treatment appraisal by subject increased covariate balance between the treatment and control groups and precision in the treatment effect estimation to a greater extent than treatment appraisal by cluster. 3) Matching retained fewer subjects and clusters than weighting, but generated greater covariate balance especially when paired with treatment appraisal by subject. 4) Aggregated covariates could reasonably replace missing true confounders in a PS analysis for estimating a treatment effect only when the correlations between the true and aggregated covariates were high ($r = .92 - .98$).

In RCTs, the random assignment mechanism often delineates the level at which treatment exposure is appraised. PS methods, which aim to mimic RCTs in nonrandomized settings, lack such clarity. The WWC (2020) standards posit that treatment should be appraised at the highest level at which all units belong to the same condition. At first glance, it then appeared contradictory that the treatment investigated by Belfi and colleagues (2016) was a characteristic of clusters (school socioeconomic composition), but they defined the probability of treatment as a property of students. All

students within a school received the same treatment, so the WWC (2020) standard implied the authors should have estimated the PS by school rather than by student. Nonetheless, the simulation results here support Belfi et al.'s (2016) approach. In contrast, Kelcey's (2011) investigation of teacher reading knowledge on student reading comprehension aligned with the WWC standard. Yet, the results from the present simulation suggest covariate balance may have improved if the PS was estimated for each student rather than each teacher.

When appraising a cluster-level treatment by subjects rather than by clusters, another peculiarity arises in scenarios where the baseline covariates are measured in the same set of clusters where treatment exposure occurs (e.g., Belfi et al., 2016; Wyse et al., 2008).[5] In this scenario, the PS model must be a single-level model that does not explicitly account for cluster dependency via fixed or random effects. The multilevel modeling literature is clear that failing to account for cluster dependency, should it exist, leads to mis-estimation of standard errors (Feller & Gelman, 2015; Raudenbush & Bryk, 2002). Yet, there is no clear evidence from simulations with clustered data in the PS literature that ignoring the cluster dependency in the PS model adversely affects the variance of the treatment effect estimation. Some studies have found multilevel models to perform better than more parsimonious models, but the improvement was descriptively small (Leite et al., 2015; Thoemmes & West, 2011). Bellara (2013) provided evidence a single-level PS model could produce superior covariate balance and treatment effect estimation than more complex models. In a simulation study, Li et al. (2013) specifically

[5] This is in contrast to studies such as Xiang and Tarasawa (2015) where subjects were nested in one set of clusters when baseline covariates were measured, but nested in a different set of clusters when treatment exposure occurred.

examined the impact of adjusting for cluster dependency in the PS model, outcome model, both models, or neither model. Their results suggested that as long as the cluster dependency was accounted for in the outcome model, the specification of the PS model had little impact on the treatment effect estimation. Consequently, for applied researchers investigating the impact of a cluster-level treatment on a subject-level outcome, the outcome model needs to account for any cluster dependency, but the PS model can be a single-level model containing both subject and cluster covariates.

Results from the present study also found that fitting the multilevel outcome model directly to the unadjusted baseline sample yielded similar treatment effect estimation as the PS methods. This aligns with the results from Leyrat and colleagues (2013) who also determined a more traditional regression approach produced smaller standard errors of the estimates. They note that a PS analysis is a two-step estimation process; first, the PS is estimated followed by estimation of the treatment effect on the outcome. The uncertainty in the first step (as demonstrated in Chapter 2) gets carried into the second estimation, thus resulting in larger standard errors than a one-step process like traditional regression methods. The regression approach also utilizes the full sample. Conversely, the results here showed that matched samples tend to be smaller, which leads to less precision.

Why then should applied researchers bother with the complexities of a PS approach? Although traditional regression methods can estimate the treatment effect while adjusting for covariates, they do not provide a framework for causal inferencing (Rubin, 2001). PS methods are based on the potential outcomes framework (Holland, 1986; Rubin, 1974). This framework posits that each subject has a unique potential

outcome resulting from each possible treatment assignment. The issue for causal inferencing is that only one of the outcomes is actually observed. When introducing PS methods, Rosenbaum and Rubin (1983) demonstrated that this issue is overcome when the average treatment effect is estimated across all subjects and the treatment and control groups are balanced on the baseline covariates related to both treatment exposure and the outcome (i.e., true confounders). Thus, PS methods can evaluate selection bias and enable stronger causal inferences. A regression model alone provides no such evidence. That being said, in the context of cluster-level treatment exposure, treatment appraisal by subject with the matching method was the only condition in the present simulation that yielded sufficient covariate balance to support causal inferencing.

Utility of Aggregated Covariates

Employing aggregated covariates as proxies for missing true cluster confounders impacted both covariate balance and treatment effect estimation. When the aggregated covariates were weakly correlated with their true counterparts (r = .23 - .44), covariate balance improved slightly, but treatment effect estimation was extremely biased. The aggregated covariates only became viable proxies in the PS analysis when they were very highly correlated (r = .92 - .98) with the missing true confounders. One caveat is that the present simulation investigated replacing 10 missing true cluster confounders with 10 aggregated covariates. Thus, any inaccuracy in the estimation resulting from weakly correlated aggregated covariates was compounded tenfold. If only replacing a single missing true confounder, it is possible the quality of the aggregated covariate could be suboptimal (e.g., r = .62 - .85) and still produce adequate treatment effect estimation.

Of course, in practice if the true confounder is missing, the correlation with the aggregated covariate is most likely unknown. Therefore, there should be strong theoretical and empirical evidence for using an aggregated covariate as a replacement for a missing true cluster confounder. For example, the percent of students receiving free/reduced-price lunch is commonly used as a proxy school-level socioeconomic status (Harwell & LeBeau, 2010). Conceptually, free/reduced-price lunch eligibility is based on the federal poverty guidelines with poverty being a narrow definition of the more multifaceted notion of socioeconomic status. Empirically, administration of the free/reduced-price lunch program leads the percent of students in a school eligible for the program to not only overestimate the true poverty rate on average, but do so inconsistently from school to school (Snyder & Musu-Gillette, 2015). Thus, despite the prima facie similarities, the theoretical and empirical evidence for using student-level free/reduced-price lunch eligibility aggregated by school as a replacement for school-level socioeconomic status in a PS analysis is insufficient given the findings from the present simulation.

To explore the utility of aggregated covariates, the simulation design assumed the true confounder was missing for all clusters, but there was complete data for the aggregated cluster value. Consider if only some clusters are missing the true confounder and an aggregated proxy covariate is also available. For instance, the true confounder is the proportion of students receiving special education services in a school, but values are missing for some schools. Yet, there is a student-level indicator for an Individualized Education Program (or IEP) for students within the schools with missing data. Can the aggregated cluster value be used to replace the missing value? Evidence from the present

simulation suggests the answer depends on the correlation between the true cluster confounder and the aggregated covariate. In this instance the correlation can be determined. An intriguing line of future research could compare the efficacy of aggregated covariates for missing cluster data to other missing data methods, such as listwise deletion and multiple imputation. In PS contexts with non-clustered data, multiple imputation has been shown to outperform other missing data methods for balancing covariates between treatment and control groups and estimation of the treatment effect (Cham & West, 2016; Leyrat et al., 2019). However, many questions remain in the development of multiple imputation for clustered data (e.g., Enders et al., 2016), including how to integrate the techniques with PS methods.

Limitations

A number of simulation design features limits the generalizability of the results from the present study. First, the simulation mimicked scenarios where baseline covariates are measured in the same set of clusters where treatment exposure occurs. In Xiang and Tarasawa (2015), for example, the baseline covariates were measured when students were in elementary school, but treatment exposure occurred in middle school. Thus, the PS model was a multilevel model with random effects for elementary school while the outcome model was a multilevel model with random effects for middle school. From the medical literature, Leyrat and colleagues (2013) explored another scenario with patients clustered within doctors. Their PS simulation mimicked a cluster RCT where doctors are randomly assigned to the treatment or control group, but selection bias arises when recruiting patients to opt in to the study. Consequently, Leyrat et al. (2013) used a true PS model that included both doctor (cluster) characteristics and patient (subject)

characteristics despite the treatment exposure being a cluster-level property. This scenario might arise in education if, for instance, a diagnostic reading test is added to schools' testing schedule, but students (or their parents) opt out of taking the test. The true PS in the present study, however, only used cluster covariates under the assumption all subjects within cluster received the treatment or control condition.

Additionally, I only compared the effectiveness of one matching and one weighting method. There are, however, numerous other matching and weighting procedures from which applied researchers could choose (Stuart, 2010). Furthermore, PS studies with clustered data have also used stratification (e.g., Kelcey, 2011; Wyse et al., 2008) and covariate adjustment on the PS (e.g., Tanner-Smith & Fisher, 2016). In a PS simulation with cluster-level treatment exposure, Yu (2012) found covariate adjustment yielded lower estimation bias than stratification. Nonetheless, comparison of conditioning methods has typically been a secondary or tertiary goal in PS simulation studies with clustered data. Prioritizing the comparison of conditioning methods across diverse contexts in the presence of clustered data would be a valuable contribution to the PS literature.

There are also a myriad of decision points and contexts in a PS analysis that could be explored further when treatment exposure is at the cluster-level, but subject-level outcomes are of interest. For instance, the true treatment effect was constrained to be homogenous across all subjects. In reality, a treatment may differentially affect certain groups of subjects. Future research could focus on expanding PS methods to detect and estimate heterogeneous treatment effects in the presence of a cluster-level treatment. Doing so would provide unprecedented nuance of the impact of school and district

changes to policy and curriculum on students. Relatedly, findings from the present study may differ with a substantially larger or smaller true treatment effect.

Conclusion

In the realm of education, changes in policy and instruction are typically implemented by classroom, school, or district rather than by individual student. Nonetheless, the primary interest is often how the changes impact student outcomes. In an RCT, the random assignment mechanism provides a clear indicator of whether the treatment and control groups were created by randomly assigning students or clusters of students. In nonexperimental studies, which are defined specifically by the lack of random assignment, the decision to appraise treatment exposure as a property of clusters or subjects is not always clear. Findings from the present study suggest that specifying the PS model and conditioning on the PS by subjects, rather than clusters, produces superior covariate balance, and thus greater reduction in selection bias. Nonetheless, additional research is needed to provide corroborating evidence and further clarity for decision-making in applied settings. Results also indicated that when a true cluster confounder is missing from a PS analysis, an aggregated covariate can be a viable replacement. There are, however, many unanswered questions regarding the quantity of, quality of, and contextual conditions under which aggregated covariates improve treatment effect estimation.

CHAPTER 4

Empirical Study - School Resource Officers and Student Social-Emotional and Academic Outcomes

Background

School security measures have markedly increased over the last two decades. The School Crime Supplement to the National Crime Victimization Survey asked a nationally representative sample of 12 – 18-year-old students whether their school uses a variety of security measures. From 1999 to 2017, students reported a rise in the presence of security guards or police officers (54% to 71% of students), locking entrance and exit doors (38% to 79%) and use of security cameras (39% to 84%; Musu, et al., 2019; Robers et al., 2013). Addington (2009) argues the proliferation of information on high profile shootings via the rising presence of the Internet and 24-hour news networks ignited parental fears of unsafe schools. This pushed governments to provide more funding for security and, in particular, school resource officers (SROs). According to the National Association of School Resource Officers (2020), an SRO is a "career law enforcement officer with sworn authority" who is intended to be a resource to school staff to resolve problems and build positive relationships with students in order to provide a safe learning environment. In contrast to security guards who are employed by school districts, SROs are uniformed police officers who are typically armed. Besides assuaging fear, there is also legitimate concern for continued security at schools. Using media reports, the Everytown for Gun Safety (2020) database documented at least 550 incidents of gunfire on school or university campuses from 2013 – 2019. Given the increasing presence of SROs, high quality research on the effects of SROs on the school environment is essential for making

informed policy decisions, including allocating funds to the most efficient security measures. Utilizing student and school information from a statewide data collection and methods for minimizing selection bias, the present study sought to examine the association between SRO presence in a school and indicators of students' social-emotional well-being and academic performance.

Visible security, including SROs, operates under the deterrence hypothesis (Tanner-Smith & Fisher, 2016; Theriot & Cuellar, 2016). The deterrence hypothesis posits that students, or anyone entering school grounds, will rationally infer that visible security measures increase their risk of being caught and punished, thereby decreasing the likelihood of committing misbehaviors and violence. By reducing misbehaviors, the goal is to increase students' perceptions of safety and support. Accordingly, when students feel safe and supported this is thought to enhance their capacity to engage in positive behaviors, especially those related to academic pursuits (Kutsyuruba et al. 2015).

Conversely, increased visible security can lead to unintended negative consequences. Students may view the presence of SROs as a show of power and control that bolsters tension rather than ameliorates fears (Bracy, 2011; Noguera, 2003). With locked doors, security cameras, and oft-armed law enforcement agents, students might become resentful of their expected passivity in the face of penological conditions. Consequently, students may then misbehave as an act of defiance. Authors also contend that some SRO actions violate students' Fourth Amendment right to secure their persons and effects from unreasonable searches and seizures without probable cause (Addington, 2009; Beger, 2003; Theriot & Cuellar, 2016). Lastly, SROs may also undermine the authority of teachers and administrators. Situations that would otherwise be handled

effectively by school staff become escalated when SROs intervene (Noguera, 2003; Weiler & Cray, 2011). Consequently, school misconduct becomes criminalized, which can lead to excessive use of force and arrests that establish a 'school-to-prison" pipeline. A holistic evaluation of SROs examining their capability to deter misbehavior without creating a negative school environment is imperative, but currently deficient.

School Security and Student Outcomes

High quality studies on the effectiveness of security measures, and SROs in particular, to reduce misbehavior and promote a safe learning environment is scant (Servoss, 2017; Weiler & Cray, 2011). From the available research, evidence that security measures successfully achieve their goals is mixed. Aggregating student self-reports to the school level, Nickerson and Martens (2008) found a positive association between overall level of school security, including SROs, with school disruption and crime. Utilizing a multilevel model, Servoss (2017) found that high security schools, 93% of which had SROs or other security personnel, reported fewer instances of misbehavior, but also lower math and reading achievement scores compared to low security schools.

SROs and Student Outcomes

Parsing the impact of a specific type of security measure compared to overall levels of security assists districts and states when making critical policy and funding decisions. Students (McDevitt & Panniello, 2005; Theriot & Orme, 2016) and principals (May et al., 2004) have largely positive views of SROs. That being said, students also express unawareness of SROs' purpose (Bosworth et al., 2011), remark that SROs make little difference on their sense of safety at school (Bracy, 2011; Theriot & Orme, 2016), and disapprove of increasing the number of SROs (Brown, 2006). These studies suggest

that SROs may not be achieving the goal of fostering productive learning environments by improving perceptions of safety. Views toward SROs may also be shifting as cases of police brutality have led to public outcry. In response to the killing of George Floyd by police officers, the third largest school district in Minnesota terminated its SRO program because an SRO presence "did not align with the priorities of the [d]istricts' equity and social emotional learning goals" (Minneapolis Board of Education, 2020).

Going beyond perceptions, Brady and colleagues (2007) explored the effectiveness of SROs to deter misbehavior when evaluating the New York City Impact Schools Initiative. The initiative doubled the number of SROs in 12 schools with the aim of reducing violent, chronically disruptive and disorderly behavior. In the first year of the initiative, researchers found Impact schools had more non-criminal police incidents, higher suspension rates, and lower attendance rates than in the year prior to the initiative. Additionally, the Impact schools had worse outcomes than 10 comparison schools. A crucial caveat, however, was the Impact schools were selected specifically because they had above average misbehavior and below average attendance. In contrast, the comparison schools were selected based only on similarity of size and racial diversity while other baseline differences were not taken into consideration.

Overall, notable gaps in each study of SROs make generalizability of their findings unclear. First, most of the studies were qualitative, which provide rich description but only from a small number of voices. For the quantitative studies, many utilized samples limited in both size and scope. Second, few studies accounted for possible contextual differences between schools. Third, most studies were observational, making them susceptible to selection bias.

Addressing Selection Bias

Randomized experiments are often deemed the gold standard for drawing causal conclusions because the randomization process, in theory, generates treatment and control groups that are similar on all observed and unobserved baseline covariates (WWC, 2020). SROs, however, are not typically randomly assigned to schools. Furthermore, SRO presence is confounded with a host of covariates that are also associated with student outcomes. For instance, high security schools have been associated with school size, proportion of minority students, students receiving free/reduced price lunch, urbanicity (Kupchik, 2010; Nickerson & Martens, 2008), lower parental education, and more students from single-parent homes (Servoss, 2017). Consequently, schools with an SRO may serve a different population of students than schools without an SRO. While regression approaches can adjust the estimated effect of an SRO by student and school characteristics, they do not establish whether the students and schools from the SRO and no SRO groups are from the same population. Thus, estimates from regression are susceptible to selection bias. Propensity score (PS) methods can mimic a randomized experiment by demonstrating that the treatment and control groups being compared are probabilistically similar on the observed baseline covariates (Austin, 2011; Rosenbaum & Rubin, 1983). PS methods can then be used in conjunction with regression to ameliorate selection bias while estimating the effect of SROs on student outcomes.

Employing PS methods, Tanner-Smith and Fisher (2016) compared schools with and without security measures. They conducted two separate analyses, one using only school-level information and the other with only student-level data. They estimated the

propensity of students and schools following one of 8 patterns of security utilization. Four patterns included the presence of security personnel who were not necessarily SROs. Results indicated that students attending schools with security personnel as the only security measure had lower grades and higher truancy rates than students in schools with no security measures or only cameras. At the school level, schools with security personnel, cameras, and metal detectors had lower average grades and attendance than other security utilization patterns. There were no differences at the school level, however, between the security personnel only group and schools with other security patterns. The quasi-experimental findings from Brady et al. (2007) and Tanner-Smith and Fisher (2016) suggest SRO presence may be associated with increased student misbehavior and truancy along with decreased academic performance.

Present Study

Despite the laudable efforts of these researchers, methodological gaps still remain in the effort to understand the degree to which SROs succeed in promoting a safe and supportive school environment where students can thrive academically and socio-emotionally. I am unaware of a study comparing schools with and without SROs while accounting for selection bias using both school and student characteristics simultaneously. Tanner-Smith and Fisher (2016) did so in separate analyses, but not simultaneously. Another distinction in the present study is the use of school covariates aggregated from student-level information. Some previous quantitative research on SROs did not consider the broader school context when evaluating the impact of SROs' presence on students (e.g., Brady et al., 2007; Theriot & Orme, 2016). When school-level

information is unavailable, covariates aggregated from student-level data can be used to account for school context.

In a systematic review of 157 primary studies, 13 non-scholarly works, and 20 other systematic reviews on school climate, Kutsyuruba and colleagues (2015) formulate the argument that a safe school environment and positive student well-being are important precursors to students' academic motivation and performance. Four student outcomes were selected for the current study to align with the stance of Kutsyuruba et al: students' sense of empowerment, perception of teacher/school support, commitment to learning, and academic performance. The measures of empowerment, which includes students' feelings of safety and being valued, and teacher/school support are most proximal to the purpose and actions of an SRO. In their review, Kutsyuruba et al. (2015) note that students feel safer in schools where their relationships with adults are respectful and caring whereas disciplinary climates lead to feelings of powerlessness, especially for minoritized students. Thus, students' sense of empowerment and perception of support are viewed as consequential for their commitment to learning and academic performance, the latter of which are the more salient goals for schools.

The purpose of the present investigation was to examine the association between SRO presence and students' school-based social-emotional well-being and academic performance using PS methods for clustered data. Within this context, however, arises the question of whether presence of an SRO should be regarded as a student or school characteristic in the analysis. In a simulation setting, results from Chapter 3 indicate that appraising SRO presence at the student level in the PS analysis should yield superior covariate balance and treatment effect estimation than appraisal at the school level. The

present study explores the two approaches in an empirical setting by conducting the analysis once with SRO as a student-level characteristic and once as a school-level characteristic.

1. What association, if any, does SRO presence have on students' sense of empowerment, perception of teacher/school support, commitment to learning, and academic performance?

2. How, if at all, does covariate balance and treatment effect estimation differ by appraising SRO presence (i.e., treatment exposure) as a school or student characteristic in the PS analysis?

Method

Sample

The study was a secondary analysis of multiple large-scale data collections. The data source for student information was the 2016 administration of the Minnesota Student Survey (MSS). The MSS is a voluntary triennial survey where students in grades 5, 8, 9 and 11 respond anonymously to a breadth of items about academics, out-of-school activities, health, violence and safety, family environment, and risky behaviors (Minnesota Department of Education [MDE], 2019a). In each administration of the survey, approximately two-thirds of all students in the selected grades and approximately half of all public and charter schools in the state participate (MDE, 2019b). The 2016 administration had a sample of 168,733 students from 1,081 schools. The primary data source for school information were the MDE data reports (MDE, 2019b), while supplemental school-level covariates were obtained by aggregating the student-level information from the MSS and from the 2015-2016 administration of the Civil Rights

Data Collection (CRDC) from the Office of Civil Rights (2019). Required of all public and charter schools, the CRDC collects information on indicators of education equity to monitor and investigate violations of civil rights laws. All three data sources were cross-sectional in nature with baseline covariates being measured at the same time as treatment exposure (i.e., SRO presence). Consequently, PS methods can reduce selection bias in the current analysis, but do not enable causal inferences between the presence of an SRO and students' school-based social-emotional well-being and academic performance.

Only students in grades 8, 9, and 11, along with the schools they attended, were included in the final analysis. Many MSS items assessing baseline characteristics relevant to the presence of an SRO and student outcomes were not asked of grade 5 students. These pertained primarily to risky behaviors, such as drug and alcohol use. For schools serving students in both grade 5 and higher grades, the grade 5 student responses were used when available to calculate the aggregated school covariates. Doing so improved the quality of the aggregated covariates by increasing the sampling ratio and reliability of the school-level values. This left a sample of 126,868 students in 553 schools.

Missing Data. Of the 553 schools, 92 (16.6%) initially had missing data on the 55 school covariates considered for the PS and outcome models. For 62 of those schools, the only missing values were from MDE and could be adequately replaced using information from the CRDC or MSS. For instance, if the percent of English Language Learner students in a particular school was missing in the MDE data, the value from the CRDC measure was used.[6] Of the remaining 30 schools (5.4% of schools), 21 were charter,

[6] The correlation between the MDE and CRDC measures for percent of English Language Learners was r = .97 and percent of special education students was r = .62. The correlation between the MDE report and MSS aggregated covariate for percent of students receiving free/reduced priced lunch was r = .93.

special education, or tribal schools whereas only 9 were traditional public schools. After removing these 30 schools, a sample of 124,997 students remained.

On 42 student-level covariates and four outcome variables, a total of 26,474 (21.1%) students had missing data. Students with missing data attended schools with SROs at a similar rate to those with complete data (87% to 82%, respectively); however, students with missing data were consistently lower on the outcome measures of empowerment (d = -0.19), teacher/school support (d = -0.14), commitment to learning (d = -0.18), and academic achievement (d = -0.30). Students with missing data were also more often non-White (42% to 27%), receiving free/reduced priced lunch (38% to 26%), and self-reported being in poor health (d = 0.14). Multiple imputation has been shown to outperform listwise deletion under the missing at random assumption in PS analyses with non-clustered data (Granger et al., 2019; Leyrat et al., 2019). With clustered data, multiple imputation techniques have been developed, but are still in their infancy with many unresolved questions (e.g., Audigier et al., 2018; Enders et al., 2016), including integration with PS methods. Therefore, missing data was handled with listwise deletion here. Given the characteristics of schools and students with missing data, the results from this study may not generalize to schools serving special populations and marginalized students with lower connections to school. The PS analysis included a final sample of 98,503 students from 523 schools. Descriptive statistics for the final sample are provided in Appendix C.

Measures

Treatment Exposure. The dichotomous treatment exposure was defined as whether or not an SRO or similar sworn law enforcement officer was present in a school.

On the MSS, students were asked "Is there a police officer or School Resource Officer (SRO) at your school?" Students could respond *Yes*, *No*, or *I don't know*. Responses of *I don't know* were coded as missing. From the CRDC, schools reported the number of full-time equivalent sworn law enforcement officers present at school or school events regardless of the source of funding for the officer(s). Schools were given a code of 1, indicating the presence of an SRO, in one of two ways. First, if schools reported a full-time equivalency > .2 in the CRDC. This corresponds to one officer being present at the school at least one day a week. This was true for 208 (39.8%) of the 523 schools (Table 4.1). Second, if at least 50% of students in a school responded *Yes* to the MSS item. Thus, for this criterion I am making the assumption that students in grade 8 or higher were able to distinguish between an SRO in law enforcement uniform and other school security personnel. The 50% threshold reflects the presence of an SRO was noticed by the majority of survey respondents within a school. An additional 126 (24.1%) schools met the MSS criterion, which brought the frequency of schools with and without an SRO to 334 (63.9%) and 189 (36.1%), respectively.

Table 4.1

Frequency of 523 Schools Meeting Criteria for School Resource Officer Designation

		CRDC	
		No	Yes
MSS	No	189	8
	Yes	126	200

Note. CRDC = Civil Rights Data Collection; MSS = Minnesota Student Survey

Student Outcomes. Measures of empowerment, teacher/school support, and commitment to learning were derived from items on the MSS. The empowerment scale

was developed as part of the Developmental Assets Profile from Search Institute (2016) and comprised of six items pertaining to feeling safe and valued. These included, “I feel safe at school” and “I am given useful roles and responsibilities.” Teacher/school support was measured with six items regarding the sense of fairness and caring students feel from teachers and school personnel. Example items include, “Overall, adults at my school treat students fairly” and “Most teachers at my school are interested in me as a person”. Commitment to learning was a 6-item measure of academic related behaviors and importance of learning, such as, “How often do you pay attention in class?” and “I think things I learn in school are useful.” Each measure was Rasch scaled with resulting values transformed so scores ranged from approximately 5 to 15 with a score of 10 indicating the midpoint of the scale. Higher values signified greater sense of empowerment, teacher/school support, or commitment to learning (see Rodriguez, 2017 for full details of construction and psychometric evaluation of these measures).

To measure academic achievement, one item on the MSS asked students, “How would you describe your grades this year?” Although typically overestimated, evidence suggests that self-reported grades are strongly correlated with actual grades (Kuncel et al., 2005; Shaw & Mattern, 2009; Sticca et al., 2017). Student responses of *Mostly As*, *Mostly Bs*, *Mostly Cs*, and *Mostly Ds* were given numeric codes of 4, 3, 2, and 1 respectively. Responses of *Mostly Fs*, and *Mostly Incompletes* were coded as 0. *None of these letter grades* was coded as missing data.

Analytic Procedure

Covariate Selection. The first step in the PS analysis was selection of student and school characteristics to include in the PS estimation model. The data from each source

(MSS, MDE, CRDC) was from a single point in time rather than prior to the introduction of an SRO in a school (i.e., treatment exposure). Consequently, covariates that could have been impacted by an SRO (i.e., endogenous variables) were removed from consideration, such as suspension rates (Kainz et al., 2017). Student covariates were derived from responses on the MSS and included demographic variables, participation in out-of-school time activities, mental and physical health, drug and alcohol use, postsecondary aspiration, and measures of social-emotional wellbeing indirectly related to school contexts. The full list is included in the Appendix along with baseline descriptive statistics. Most student covariates were dichotomized as being true or not in the last 30 days. Continuous measures of being bullied, mental distress, social competence, and family/community support were constructed following the same process as the outcomes (e.g., empowerment). The process is outlined in Rodriguez (2017).

As reported by MDE or collected in the CRDC, school covariates included grade levels served, charter/magnet designation, urbanicity, student composition, academic performance, and presence of support services staff. School covariates also included those aggregated from student responses.[7] Some covariates were reported from multiple sources. MDE, CRDC, and MSS all had measures of school size for example. When this situation arose, only one source was used in the PS and outcome models. When available, the information from MDE was prioritized followed by CRDC, and lastly the MSS.

[7] All of the aggregated covariates were largely formative in nature with a sampling ratio between. .33 and .37. As noted in Chapter 2, it is difficult in practice to determine whether a low sampling ratio is an indication of poor measurement of school values. As a sensitivity check, the PS analysis was also run without the use of aggregated covariates. Compared to the results presented in the paper, the sensitivity analysis had slightly worse covariate balance and treatment effect estimates had larger standard errors, but did not differ substantially in magnitude.

PS Model Specification. Two PS estimation models were compared, which aligned with those examined in Chapter 3. The first – the Cluster Level (CL) model – assessed presence of an SRO ($Z_j = 1$) as a property of schools:

$$\text{logit}(Z_j = 1) = \gamma_0 + \sum_{q=1}^{55} \gamma_q W_{qj} \quad (17)$$

The single-level model included 55 school-level covariates, W_q, to estimate a unique PS for each school *j* with every student within a school having the same PS. The other model – the Subject Level (SL) model – defined the PS as the probability of a student attending a school with an SRO:

$$\text{logit}(Z_i = 1) = \gamma_0 + \sum_{p=1}^{42} \gamma_p X_{pi} + \sum_{q=43}^{55} \gamma_q W_{qi} \quad (18)$$

This single-level model produced a PS for each student *i* based on 42 student characteristics, X_{pi}, and 55 school characteristics, W_{qj}. The school characteristics were treated as student-level covariates, so the *j* subscript for schools was excluded. Results from Chapter 3 indicated that the CL model (Equation 17) and SL model (Equation 18) produced similar treatment effect estimation; however, the SL model yielded superior covariate balance, indicating greater reduction in selection bias.

Conditioning Method. Following estimation of the PS, the SRO (treatment) and No SRO (control) groups were equated by school for the CL model and by student for the SL model. The matching method that outperformed the weighting method for covariate balance in Chapter 3 was utilized here. As one of the more commonly employed matching methods (Thoemmes & Kim, 2011; Zakrison et al., 2018), a school (or student) with an SRO was paired with a school (or student) without an SRO within a caliper width of 0.2 standard deviations of the logit of the PS on a 1:1 basis without replacement using a nearest neighbor greedy matching algorithm.

Evaluating Covariance Balance. Covariate balance between the SRO and No SRO groups was then assessed on both student and school covariates at the student level, which aligns with WWC (2020) standards. To address the second research question, covariate balance was assessed for both the CL and SL sample, as well as for the full unadjusted sample. As described in Chapter 3, an absolute standardized mean difference < 0.10 and a variance ratio < 2.0 are generally accepted as evidence of balance between the SRO and No SRO groups (Austin, 2011; Rubin, 2001).

Outcome Analysis. The primary research question asked whether students' sense of empowerment, teacher/school support, commitment to learning, and academic performance differed based on the presence or absence of an SRO. A descriptive comparison was conducted first. The mean and standard deviation of each outcome was calculated for the SRO and No SRO groups as well as the standardized mean difference between the two groups. The comparisons were calculated on the sample produced from each of the two PS model specifications (CL and SL) and the full unadjusted sample.

To account for the clustering of students within school, a multilevel model was fit for each of the four outcomes. WWC (2020) standards require any covariates with an absolute standardized difference > 0.05 between the SRO and No SRO group be included in the outcome model. Ho and colleagues (2007) further argue that PS methods pertain to study design, so the outcome model used for analysis should be the same regardless of whether PS methods were employed. Therefore, the multilevel model included the 42 student covariates and 55 school covariates in addition to the indicator for presence of an SRO and a random intercept for school. The outcome model is represented as:

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{42} \gamma_{p0} X_{pij} + \sum_{q=1}^{55} \gamma_{0q} W_{qj} + \gamma_{0\,56} Z_j + u_{0j} + r_{ij} \quad (19)$$

$u_{0j} \sim N(0, \tau_{00})$; $r_{ij} \sim N(0, \sigma^2)$

where Y_{ij} is the outcome for student i in school j, and u_{0j} and r_{ij} are the random intercept for school and student-level error term, respectively[8]. Once again, the process was repeated on CL sample, SL sample, and full unadjusted sample. In the CL sample and full unadjusted sample, the SRO indicator Z was appraised as a school characteristic, which is the representation shown in Equation 19. In the SL sample, Z was conceptually a student characteristic, which would instead be represented in the model as $\gamma_{43\,0}Z_{ij}$. Nonetheless, estimation of the parameter γ for Z is equivalent in both conceptual representations.

In PS parlance, given that a matching method was used to condition the SRO and No SRO groups on the PS, the multilevel model estimated the average treatment effect for the treated. Thus, the regression estimate is interpreted as the expected change in the outcome for a student moving from the No SRO group to the SRO group. This interpretation implies causality. However, the data for this study were cross-sectional, which does not enable causal inferences to be drawn. A more appropriate interpretation for the resulting estimate is the association between presence of an SRO and the student outcomes while controlling for the other covariates in the model, adjusting standard errors for cluster dependency, and reducing selection bias. The ability to quantify and adjust for selection bias is what differentiates PS methods from fitting the multilevel model to the full unadjusted sample.

In addition to evaluating the magnitude and precision of the regression estimate from each model, the f^2 value of the SRO indicator was calculated. As an effect size for a

[8] Model assumptions were checked using QQ plots of the u_{0j} and r_{ij} values, scatterplot of u_{0j} and r_{ij} values, and scatterplot of r_{ij} and fitted Y_{ij} values.

fixed effect, f^2 is interpreted as the proportion of variance in the outcome explained by the SRO indicator relative to the proportion of unexplained outcome variance. Using Equations 4 and 5 from Lorah (2018), calculating f^2 also necessitated running the multilevel model from Equation 19 once while excluding the SRO indicator and once as an intercept-only model (i.e., no covariates nor SRO indicator) for each outcome.

Results

Covariate Balance

In the full unadjusted sample, an SRO was present in 63.9% of schools, which were attended by 82.2% of students (Table 4.2). Matching by student on a 1:1 basis constrained the proportion of students attending a school with and without an SRO to be equal in the SL sample; however, the proportion of schools differed. Conversely, in the CL sample where schools were matched on a 1:1 basis, the proportion of schools with or without an SRO was equal, but the proportion of students differed. The CL sample also retained fewer total students (n = 29,592) and schools (n = 232) than the SL sample (students = 30,464; schools = 436). For reference, the SL sample retained 30.9% of students from 83.3 % of the schools in the full sample.

Table 4.2

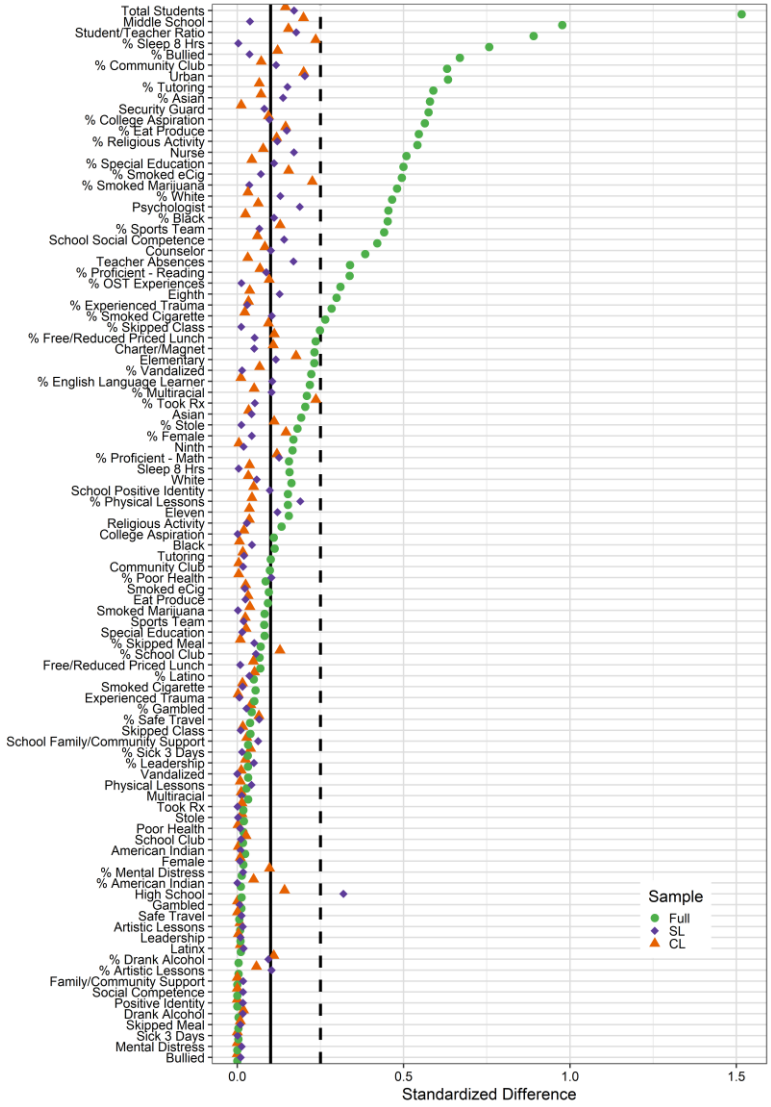
Frequency (Percent) of Size and Balanced Covariates

	Students		Schools		Absolute	Variance
	SRO	No SRO	SRO	No SRO	Std. Diff	Ratio
Full	80926 (82.2)	17577 (17.8)	334 (63.9)	189 (36.1)	46 (47.4)	43 (82.7)
SL	15232 (50.0)	15232 (50.0)	251 (57.6)	185 (42.4)	70 (72.2)	50 (96.2)
CL	16262 (55.0)	13330 (45.0)	116 (50.0)	116 (50.0)	76 (78.4)	50 (96.2)

Note. SRO = School Resource Officer; SL= Sample matched by student; CL = Sample matched by school (i.e., cluster)

Figure 4.1

Absolute Standardized Mean Difference Between SRO and No SRO Groups in the Full and Propensity Score Samples



Note. Solid line indicates standardized difference of 0.10 with the dashed line at 0.25. SL= Sample matched by student; CL = Sample matched by school (i.e., cluster).

Both the CL and SL samples reduced selection bias compared to the full sample by improving the covariate balance between the SRO and No SRO groups. That being

said, selection bias was not eliminated as some covariate imbalance remained in both PS samples. A summary of the balance for each sample is provided in Table 4.2 and the absolute standardized difference for individual covariates is displayed in Figure 4.1.

The variance ratio between the SRO and No SRO groups was balanced (< 2.0) for only 43 (83%) of the 52 continuous covariates in the full sample. Both the CL and SL samples had balance on 50 (96%) of the 52 continuous covariates. The absolute standardized difference between the SRO and No SRO groups was < 0.10 for only 46 (47%) of the 97 total student and school covariates in the full sample. The balance improved in the SL sample (70 covariates, 72%) and increased further in the CL sample (76 covariates; 78%). These results are contradictory to the simulation findings in Chapter 3 where SL samples were consistently more balanced. Although the CL and SL samples had better balance than the full sample, the SRO and No SRO groups were still unbalanced on over 20% of the covariates. Consequently, selection bias was reduced, but not eliminated.

SRO Presence and Student Outcomes

A descriptive comparison of the student outcomes found little difference between the SRO and No SRO students (Table 4.3). Using the CL sample for example, students in the SRO group reported a mean sense of empowerment of 12.45 (*SD* = 1.93). This indicates that the students, on average, felt safe and appreciated rather than unsafe and unappreciated. Students in the No SRO group felt similarly (*M* = 12.43; *SD* = 1.93), producing a negligible standardized mean difference (*d* = 0.01). An analogous pattern between the SRO and No SRO students manifested across the other outcomes with students in both groups, on average, reporting they felt more supported adults in the

school than not, were committed to learning, and received mostly B grades. The findings were also comparable in the SL sample and full sample. Although still small, the full sample consistently had slightly higher standardized mean differences, which could be an indication of selection bias.

Table 4.3

Mean (Standard Deviation) and Standardized Mean Difference in Student Outcomes

	Empowerment			Teacher/School Support		
Sample	SRO	No SRO	*d*	SRO	No SRO	*d*
Full	12.49 (1.92)	12.43 (1.92)	0.03	11.78 (2.21)	11.73 (2.28)	0.03
SL	12.40 (1.93)	12.43 (1.93)	-0.02	11.70 (2.26)	11.72 (2.29)	-0.01
CL	12.45 (1.93)	12.43 (1.93)	0.01	11.73 (2.26)	11.69 (2.30)	0.02
	Commitment to Learning			Grade Point Average		
Sample	SRO	No SRO	*d*	SRO	No SRO	*d*
Full	12.19 (1.53)	12.08 (1.55)	0.07	3.22 (0.91)	3.16 (0.94)	0.06
SL	12.12 (1.53)	12.10 (1.56)	0.01	3.15 (0.95)	3.17 (0.94)	-0.02
CL	12.12 (1.54)	12.10 (1.55)	0.01	3.17 (0.95)	3.17 (0.95)	0.00

Note. SL= Sample matched by student; CL = Sample matched by school (i.e., cluster); SRO = School Resource Officer

Results from the multilevel outcome models concurred with the descriptive comparison. When controlling for the other covariates in the model and adjusting standard errors for cluster dependency, there was no association between the presence of an SRO and the student outcomes (Table 4.4). Additionally, the unstandardized regression estimate, range of the 95% confidence interval, and f^2 estimate, were comparable across both the SL and CL samples, as well as the full sample.

Table 4.4

Multilevel Regression Estimate and Variance Explained for School Resource Officer Presence on Student Outcomes

	Empowerment			Teacher/School Support		
Sample	*b*	95% CI	f^2	*b*	95% CI	f^2
Full	-0.02	[-0.05, 0.01]	.00	-0.02	[-0.10, 0.06]	.00
SL	-0.02	[-0.06, 0.01]	.00	-0.01	[-0.09, 0.07]	.00
CL	-0.01	[-0.05, 0.02]	.00	-0.03	[-0.12, 0.07]	.00
	Commitment to Learning			Grade Point Average		
Sample	*b*	95% CI	f^2	*b*	95% CI	f^2
Full	0.01	[-0.03, 0.05]	.00	-0.02	[-0.05, 0.02]	.00
SL	0.00	[-0.04, 0.05]	.00	-0.01	[-0.05, 0.03]	.00
CL	-0.01	[-0.06, 0.04]	.00	-0.02	[-0.06, 0.02]	.00

Note. b = Unstandardized estimate while controlling for student and school covariates; SL= Sample matched by student; CL = Sample matched by school (i.e., cluster)

Discussion

The substantive purpose of this study was to explore the association between the presence of an SRO in a school and students' school-based social-emotional well-being and academic performance. The National Association of School Resource Officers (2020) states the role of an SRO is to be a resource to school administrators and build positive relationships with students in order to provide a safe learning environment. In their systematic literature review of school climate, Kutsyuruba and colleagues (2015) posit that when students feel safe and supported in their school environment, they have greater capacity for academic pursuits which can lead to improved performance. However, the present study did not find evidence of SRO presence being associated with this process. Specifically, results showed no association between SRO presence in a school and students' sense of empowerment, perceptions of teacher/school support,

commitment to learning, and self-reported academic performance. These findings provide a notable contribution to the literature on SRO presence in school. The extant research was predominantly qualitative studies or based on samples with limited generalizability. The present study broadened the scope of the findings by utilizing student information from a statewide data collection and school information reported by state and federal agencies. Additionally, most previous quantitative work was susceptible to selection bias. Although selection bias was not eliminated, the PS methods employed here reduced selection bias and demonstrated the extent to which schools with an SRO served a comparable population of students as the schools without an SRO.

Results from the current analysis also largely aligned with the limited previous research. Using survey responses from one school district, Theriot and Orme (2016) found that direct interaction with an SRO and even taking a course led by an SRO had no association with feeling safe at school. Compared to students who felt unsafe, students who felt safe did report more positive attitudes toward SROs, as well as greater school connection. Thus, the authors conjectured that the positive attitudes toward SROs might be a reflection of students' general feeling toward school rather than a specific relationship between students and SROs. Interviews with high school students also revealed that SRO presence had little bearing on their sense of safety (Bracy, 2011). Students typically felt their school was already safe and there was little need for an SRO or they believed those who would commit a crime would be undeterred by an SRO.

Despite falling directly under the purview of an SRO's purpose in schools, research thus far has found little to no association between SRO presence and students' feeling of safety, empowerment, and support. Therefore, it is unsurprising there was a

lack of association with the more ancillary commitment to learning and academic performance outcomes in the present study. Using national datasets and PS methods, Tanner-Smith and Fisher (2016) did find the presence of security personnel was associated with lower student academic performance compared to other visible security measures. Theriot and Cuellar (2016) note, however, that it is important to distinguish SROs from other security personnel. Compared to security guards and other school personnel, SROs have sworn authority to conduct searches, make arrests, and share information with juvenile courts. Consequently, authors argue that students' rights can be jeopardized (Beger, 2003; Noguera, 2003; Weiler & Cray, 2011). This led to the United States Department of Education (2014) adopting guidelines that SROs should receive specialized training on child and adolescent development and psychology, age-appropriate behavioral interventions, and restorative justice techniques. The specialized training SROs undergo distinguishes them from other law enforcement agents. Thus, the present study extended Tanner-Smith and Fisher's (2016) work by looking at SROs specifically rather than security personnel and security measures more generally.

The analysis conducted here also aimed to compliment Tanner-Smith and Fisher (2016) from a methodological stand point by further developing the use of PS methods to reduce selection bias. Analyzing the student and school levels separately, they found that security personnel alone had little association at the school level, but was associated with lower academic performance at the student level. In comparison, I used a multilevel outcome model to incorporate student and school level information simultaneously and found no association between SRO presence and student academic performance. For estimating the PS, the PS literature was unclear as to whether SRO presence should be

appraised as a student or school-level property. Results from Chapter 3 suggested treating SRO presence at the student level would best minimize selection bias. Nonetheless, the analysis was run using both approaches as well as once without any PS methods. Both PS analyses reduced selection bias, but unlike the findings from Chapter 3, estimating the PS and matching by school ameliorated selection bias to a greater extent than estimating the PS and matching by student. That being said, the SRO and No SRO groups remained unbalanced on some covariates indicating that selection bias was not eliminated. Along with the cross-sectional nature of the data, the lingering selection bias means that study results evaluated a correlational, and not causal, association between SRO presence and student outcomes.

Limitations and Future Research

The utilization of large-scale data collections addressed a generalizability issue in previous research, but the nature of a secondary data analysis also meant that some assumptions were made regarding data quality. Efforts to ensure data quality included checking for values outside of expected ranges, using covariates with low rates of missing data, and comparing values for measures reported across multiple datasets. Nonetheless, as shown in Table 4.1 when constructing the SRO indicator, there were still discrepancies between data sources that imply the findings should be interpreted with some caution. Thus, there remains a need for a study that is large in scope, but also directly addresses possible effects of SRO presence rather than relying on a secondary analysis of previously collected data.

Additionally, the current analysis presented an exploration of the association between SRO presence and outcomes for a general population of students. In doing so,

the analysis estimated the association homogenously across the students. A growing body of authors argues that SROs may have a negative impact on certain subpopulations of students (e.g., Bracy, 2011; Kutsyuruba et al., 2015). Kutsyuruba and colleagues (2015) found through their review that in addition to discriminatory discipline policies by gender, sexual, and racial identities, as well as behavioral needs, students from these minoritized populations are also more often the victims of bullying and violence. Consequently, the guiding principles for school climate and discipline document from the United States Department of Education (2014) acknowledges that students of color and students with disabilities disproportionately have higher contact with school-based law enforcement. For instance, Servoss (2017) utilized a multilevel model and found a cross-level interaction between the amount of school security and the disparity in attendance and disruptiveness between Black and White students. At low security schools, there were no racial differences in teachers' reports of disruptiveness and attendance. In high security schools, Black students were more often reported as disruptive and having more attendance-related misbehavior than White students. Thus, marginalized students are not only more likely to come into contact with SROs either as perpetrators or victims, but also feel less safe and supported in their school environment (Kutsyuruba et al., 2015; Theriot & Orme, 2016).

The present study attempted to make the SRO and No SRO groups comparable by including racial and gender identity, experiences being bullied, and exposure to trauma among the covariates in the PS estimation model and in the outcome models. Nonetheless, the missing data analysis revealed that racial minoritized students and students who reported low levels of empowerment, teacher/school support, commitment

to learning, and academic performance were underrepresented in the sample. Consequently, the results may not generalize to students on the academic and social periphery of schools. Future research efforts should focus specifically on the effect of SRO presence near and interactions with students from vulnerable and marginalized populations.

Conclusion

The majority of students in the present study attended a school with an SRO – a uniformed and often armed law enforcement agent. The purpose of an SRO is to foster positive relationships with students in order to create a safe learning environment. To inform policy decisions, it is critical to understand the extent SROs' are achieving this goal. The present study contributed to this understanding while addressing methodological and generalizability gaps in the current literature. The results indicated that SRO presence in schools was not associated with students' sense of empowerment, perceptions of teacher/school support, commitment to learning, or academic performance. While this finding is indicative of a general population of middle and high school students, future research needs to focus specifically on how SRO presence may have different effects on student populations that have historically been targeted by discrimination and school violence.

CHAPTER 5

Conclusion

My dissertation explored conceptual and practical issues at the intersection of propensity score (PS) methods, clustered data, and aggregated covariates in the context of education research. The clustering of students within classrooms, schools, and districts presents study design and statistical challenges when attempting to evaluate education interventions and reforms. Ideally, the casual impact of any reform would be assessed by randomly assigning some students to a treatment condition and other students to a control condition. Changes in school policies and practices, however, are often implemented by school rather than by student. Therefore, random assignment needs to occur by school (i.e., by cluster). The example highlighted in Chapter 1 and investigated in Chapter 4 was the effect of a School Resource Officer (SRO) on students' social-emotional well-being and academic performance. Although randomly assigning schools to an SRO or no SRO group is theoretically possible, doing so is logistically untenable in practice. Therefore, investigating impacts of an SRO on the school environments must rely on nonrandomized data. However, nonrandomized studies are susceptible to selection bias. Selection bias is present when the students and schools in the treatment group systematically differ from the students and schools in the control group. Consequently, causality of the treatment (e.g., SRO presence) cannot be parsed from the possible confounding effect of the systematic differences between treatment and control groups.

PS methods were developed to mimic randomized controlled trials (RCTs) with nonrandomized data (Austin, 2011; Rosenbaum & Rubin, 1983). In their seminal paper, Rosenbaum and Rubin (1983) demonstrated that under certain assumptions, PS methods

can produce unbiased estimation of the treatment effect. However, there are two properties of random assignment in an RCT that are particularly useful with clustered data and subject-level outcomes. In order for treatment effect estimation to remain unbiased in a PS analysis, steps need to be taken to address the absence of a random assignment mechanism and the properties it possesses. The simulations in Chapters 2 and 3 were designed to better inform methodologists and applied researchers on the appropriate steps. Chapter 4 then adapted the guidance drawn from the simulation studies to a practical problem with real data.

Appraising Treatment Exposure by Subjects or Clusters

The first property informs the appraisal of treatment exposure by subjects or by clusters. In an RCT, the level at which random assignment occurs clearly delineates the level at which treatment exposure should be appraised. If students are randomly assigned to a reading intervention, then some students in a classroom receive the intervention whereas other students in the same classroom do not. If classrooms are randomly assigned, then every student within a treatment classroom receives the intervention and every student in a control classroom receives business-as-usual. Given the absence of a random assignment mechanism, the level at which treatment exposure should be appraised in a PS analysis can be vague. The ambiguity is most apparent when treatment exposure is a property of clusters. On the one hand, every subject within the cluster receives the same treatment. The WWC (2020) standard asserts such a study is a clustered design with treatment exposure appraised by clusters. This implies the PS procedures should entail estimating the PS and conditioning the treatment and control groups by clusters. On the other hand, supporting causal inferences with a subject-level

outcome necessitates demonstrating treatment and control groups are balanced on subject covariates. Doing so requires the PS be estimated for each subject and with conditioning also occurring by subject.

Covariate Balance

Treatment appraisal decisions were explored using a simulation study in Chapter 3 and empirical data in Chapter 4. The simulation was designed so the true probability of treatment (i.e., true PS) was entirely due to variation in cluster-level covariates. Yet, covariate balance between the treatment and control groups was greater when treatment exposure was appraised by subjects in the PS analysis. Thus, the simulation results suggested that selection bias was reduced to a larger extent when both subject and cluster covariates were included in the PS estimation model. That being said, acceptable covariate balance was only achieved when subject-level treatment appraisal was paired with the matching conditioning method. Selection bias remained with the weighting method and for all conditions where treatment exposure was appraised by clusters. The empirical study in Chapter 4 also demonstrated covariate balance issues. In contrast to the simulation results, balance between students in the SRO and no SRO groups was slightly better with treatment appraisal by clusters. Achieving sufficient balance between the treatment and control groups is a necessary antecedent for drawing causal inferences from a PS analysis. Therefore, applied researchers are encouraged to use the PS model and conditioning method that yields the best covariate balance (Ho et al., 2007; Kainz et al., 2017). Results from my dissertation support this guideline. The evidence from the simulation and empirical study is not strong enough to declare any combination of

treatment appraisal and conditioning method as definitively superior for balancing covariates.

Comparing covariate balance when a cluster-level treatment is appraised by subject or clusters is a promising avenue for future research. The simulation in Chapter 3 is, to my knowledge, the only simulation study that evaluated covariate balance when treatment exposure was a property of clusters in a PS analysis. Previous PS simulations with cluster-level treatment exposure only examined impacts on treatment effect estimation (Leyrat et al., 2013; Yu, 2012). Generalizability of the findings from Chapter 3 are of course limited by the constraints placed on the simulation design. Future research needs to explore whether challenges balancing treatment and control groups arise with different constraints and contexts. One possibility is that covariate balance is uniformly more difficult to achieve in PS studies with cluster-level treatment exposure rather than exposure by subjects. Even in RCTs, random assignment by clusters presents unique challenges that makes selection bias more persistent (Brierley et al., 2012).

Treatment Effect Estimation

The simulation in Chapter 3 found few differences between treatment appraisal by subjects and by clusters when estimating the treatment effect. Only with 60 clusters and paired with the matching conditioning method did treatment appraisal by cluster produce larger variance in the treatment effect estimation. The full sample in Chapter 4 consisted of 532 schools. Although matching reduced the number of schools, both PS samples contained over 200 schools. Given the number of schools (i.e., clusters), the negligible difference in the magnitude and the standard error in the treatment effect estimates

between the subject-level and cluster-level treatment appraisal aligned with the simulation findings.

Regarding treatment effect estimation accuracy, the overarching guideline for applied researchers is the importance of having a sufficient number of clusters. The simulation in Chapter 2 was designed to explore the sample characteristics generated from four procedures for creating aggregated covariates. Specifically, the sample characteristics needed to be suitable for a PS analysis of clustered data with treatment exposure being a property of clusters. The results of the initial simulation and a follow-up simulation revealed that a minimum of 60 clusters were needed for the PS model to converge with any consistency. The simulation in Chapter 3 found that with 60 clusters convergence issues in the outcome model persisted even after removing the replications where the PS model failed to converge. When both the PS and outcome models converged, variance in the treatment effect estimation was larger in the 60-cluster condition than the 100- or 140-cluster conditions.

The minimum viable sample size is an important issue for applied researchers. Increasing the number of subjects costs time and money. The findings from Chapters 2 and 3 suggest the overall number of subjects is less consequential than increasing the number of clusters, which aligns with findings from the broader multilevel modeling literature (e.g., Maas & Hox, 2005). Nonetheless, a valuable line of future research is exploring whether the minimum number of clusters needed for a PS analysis with cluster-level treatment exposure varies under different conditions. For instance, in my simulations the true PS was calculated with 10 covariates and with the PS and outcome models containing 10 or 30 covariates. The minimum number of clusters may vary

depending on how many covariates are used to generate the PS and in the estimation models.

Aggregated Covariates

The second beneficial property of random assignment pertains to covariates missing from the PS analysis. In an RCT, random assignment is expected to balance treatment and control groups on both measured and unmeasured covariates. In the presence of clustered data, random assignment should, in theory, balance the two groups on covariates at both the subject and cluster level. In contrast, PS methods can only achieve balance on the covariates included in the PS model. When a true confounder – a covariate related to both treatment exposure and the outcome – is unmeasured, selection bias between the treatment and control groups can persist (Imbens, 2004; Rosenbaum & Rubin, 1983). With both non-clustered and clustered data, the PS literature has established that omitting a true confounder from the PS analysis increases bias in treatment effect estimation (Austin et al., 2007; Kainz et al., 2017; Kelcey, 2009; Leyrat et al., 2013). Therefore, when a true confounder is unmeasured, PS methods no longer support causal inferences. Aggregated covariates have the potential to alleviate this pervasive problem. Although aggregated covariates are common in multilevel models for clustered data, little is known about their efficacy in PS analyses.

Using subject-level information aggregated to the cluster-level, the simulation in Chapter 3 determined that aggregated covariates can be a viable proxy for true cluster-level confounders missing from the PS analysis. However, the utility of the aggregated covariates to minimize bias and variance in the treatment effect estimation was highly dependent on the correlation between the aggregated covariates and the missing true

confounders. Estimation bias and variance was only acceptable with correlations of .92 - .98. With 10 true confounders being replaced by aggregated covariates, estimation bias and imprecision became concerningly large even with correlations of .62 - .85.

Replacing 10 true confounders was intended to represent typical practice in applied settings. The number of aggregated covariates commonly used in a PS analysis is currently unknown. Nonetheless, a review of non-clustered PS studies found the majority of studies included at least 15 covariates (Thoemmes & Kim, 2011). Maintaining the unconfoundedness assumption with clustered data presumably requires a larger set of covariates. A future study to compliment the Chapter 3 simulation can explore how treatment effect estimation is impacted by the number of aggregated covariates included in the PS analysis. Furthermore, the simulation did not establish how the poor treatment effect estimation when the true and aggregated covariates were correlated between .62 - .85 compared to omitting the true confounder without a replacement.

The implication from Chapter 3 for applied researchers is the inclusion of aggregated covariates in a PS model as a proxy for a true cluster confounder requires careful consideration of the theoretical and empirical evidence. In practice, however, determining whether an aggregated covariate possesses the necessary measurement qualities can be extremely challenging. First, if a true confounder is unmeasured, the correlation with aggregated covariate in the available sample will be unknown. Second, parsing the degree to which a covariate is a formative or reflective aggregation is often imprecise. Consequently, whether measures of cluster dependency (ICC(1)), reliability of cluster values (ICC(2)), or sampling ratio should be used to indicate quality of the aggregation is rather ambiguous. For instance, an aggregated covariate used in the

empirical study in Chapter 4 was the percent of students in each school who reported eating at least one fruit and one vegetable on a typical day. The survey item asked students about their own behavior, suggesting the subsequent aggregation was formative in nature. That being said, for many students, what they eat for lunch during the week is dependent on the food served at school. Thus, students' behavior may be reflective of the availability of fruits and vegetables at school. If the type of aggregation is uncertain, how should ICC(2) and sampling ratio be interpreted, if at all? A low sampling ratio is not commensurate with the sample being unrepresentative cluster population. The results from Chapter 3 also suggest that the reliability of the aggregated values has little impact on treatment effect estimation. Additional research needs to clarify whether applied researchers should concern themselves with indicators of aggregated covariate quality. If yes, then how to do so appropriately.

Reviewing the literature on the measurement quality and application of aggregated covariates also revealed multiple procedures for generating aggregated covariates in simulation studies. There was, however, no clear indication of which procedure would be most suitable in the context of a PS analysis of clustered data. Consequently, before conducting a simulation of PS methods with aggregated covariates (Chapter 3), I first needed to explore which procedure for generating aggregated covariates produced sample datasets with characteristics amenable to a PS analysis (Chapter 2). A primary finding from Chapter 2 is the four procedures generated starkly different sample characteristics. Based on their appropriateness in previous studies, all four procedures appeared viable at first, but only the Formative-RE procedure met the criteria I needed for the Chapter 3 simulation. In contrast, the samples generated from the

other three procedures did not possess characteristics representative of education data. This finding highlights the importance of checking the comparability of the generated samples to the intended real-world context. Unfortunately, Harwell and colleagues (2018) found in a review of 677 simulation studies, only 2.2% reported doing so. Although not utilized in my dissertation, a common tactic is to base the simulation designs and parameter constraints on an empirical dataset. In such cases, the efficacy of the simulation conclusions is highly dependent the generated samples sharing the characteristics of the empirical dataset.

References

Addington, L. A. (2009). Cops and cameras: Public school security as a policy response to Columbine. *American Behavioral Scientist*, *52*(10), 1426-1446. https://doi.org/10.1177/0002764209332556

Ali, M. S., Groenwold, R. H. H., Pestman, W. R., Belitser, S. V., Roes, K. C. B., Hoes, A. W., … Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology and Drug Safety*, *23*(8), 802–811. https://doi.org/10.1002/pds.3574

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

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., … Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, *33*(2), 160–183. https://doi.org/10.1214/18-STS646

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

Austin, P. C. (2009b). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, *29*(6), 661–677. https://doi.org/10.1177/0272989X09341755

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057–1069. https://doi.org/10.1002/sim.6004

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*(4), 734–753. https://doi.org/10.1002/sim.2580

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate

causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*, *9*(6), 377–385. https://doi.org/10.1111/j.1524-4733.2006.00130.x

Beal, D. J., & Dawson, J. F. (2007). On the use of Likert-type scales in multilevel data: Influence on aggregate variables. *Organizational Research Methods*, *10*(4), 657–672. https://doi.org/10.1177/1094428106295492

Beger, R. R. (2003). The" worst of both worlds": School security and the disappearing fourth amendment rights of students. *Criminal Justice Review*, *28*(2), 336-354. https://doi.org/10.1177/073401680302800208

Belfi, B., Haelermans, C., & De Fraine, B. (2016). The long-term differential achievement effects of school socioeconomic composition in primary education: A propensity score matching approach. *British Journal of Educational Psychology*, *86*(4), 501–525. https://doi.org/10.1111/bjep.12120

Bellara, A. P. (2013). *Effectiveness of propensity score methods in a multilevel framework: A Monte Carlo study* [Doctoral dissertation, University of South Florida]. https://scholarcommons.usf.edu/etd/4635/.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, *1*(4), 355–373. https://doi.org/10.1177/109442819814001

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 349-381). San Francisco, CA: Jossey-Bass.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods*, *10*(4), 551–563. https://doi.org/10.1177/1094428107301102

Bloom, H. S., Bos, M., & Lee, S.-W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445–469. https://doi.org/10.1177/0193841X9902300405

Bosworth, K., Ford, L., & Hernandaz, D. (2011). School climate factors contributing to student and faculty perceptions of safety in select Arizona schools. *Journal of School Health, 81*, 194–201. https://doi.org/10.1111/j.1746-1561.2010.00579.x

Bracy, N. L. (2011). Student perceptions of high-security school environments. *Youth & Society*, *43*(1), 365-395. https://doi.org/10.1177/0044118X10365082

Brady, K. P., Balmer, S., & Phenix, D. (2007). School-police partnership effectiveness in urban schools: An analysis of New York City's impact schools initiative. *Education and Urban Society*, *39*(4), 455-478. https://doi.org/10.1177/0013124507302396

Brierley, G., Brabyn, S., Torgerson, D., & Watson, J. (2012). Bias in recruitment to cluster randomized trials: a review of recent publications. *Journal of Evaluation in Clinical Practice, 18*(4), 878–886. https://doi.org/10.1111/j.1365-2753.2011.01700.x

Brown, B. (2006). Controlling crime and delinquency in the schools: An exploratory study of student perceptions of school security measures. *Journal of School Violence, 4*(4), 105–125. https://doi.org/10.1300/J202v04n04_07

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

Cham, H., & West, S. (2016). Propensity score analysis with missing data. *Psychological Methods, 21*(3), 427–445. https://doi.org/10.1037/met0000076

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265–2281. https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222-240. http://dx.doi.org/10.1037/met0000063

Everytown for Gun Safety (2020). *Gunfire on school grounds in the United States.* Retrieved May 26, 2020 from https://everytownresearch.org/gunfire-in-school/.

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009–2015. *Educational Researcher*, *47*(4), 221–234. https://doi.org/10.3102/0013189X18759524

Feller, A., & Gelman, A. (2015). Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences* (pp. 1–16). https://doi.org/10.1002/9781118900772.etrds0160

Goldfeld, K. (2020). *simstudy: Simulation of study data* [R package].

Granger, E., Sergeant, J. C., & Lunt, M. (2019). Avoiding pitfalls when combining multiple imputation and propensity scores. *Statistics in Medicine*, *38*(26), 5120–5132. https://doi.org/10.1002/sim.8355

Greifer (2020). *WeightIt: Weighting for covariate balance in observational studies* [R package].

Griswold, M. E., Localio, A. R., & Mulrow, C. (2010). Propensity score adjustment with multilevel data: Setting your sites on decreasing selection bias. *Annals of Internal Medicine*, *152*(6), 393-395. https://doi.org/10.7326/0003-4819-152-6-201003160-00010

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (Vol. 11). SAGE publications.

Harwell, M., Kohli, N., & Peralta-Torres, Y. (2018). A survey of reporting practices of computer simulation studies in statistical research. *The American Statistician*, *72*(4), 321–327. https://doi.org/10.1080/00031305.2017.1342692

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher*, *39*(2), 120–131. https://doi.org/10.3102/0013189X103

Heil, S., Reisel, L., & Attewell, P. (2014). College selectivity and degree completion. *American Educational Research Journal*, *51*(5), 913–935. https://doi.org/10.3102/0002831214544298

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. https://doi.org/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.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960. https://doi.org/10.1080/01621459.1986.10478354

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. https://doi.org/10.1198/016214506000000447

Hughes, J. N., Chen, Q., Thoemmes, F., & Kwok, O. (2010). An investigation of the relationship between retention in first grade and performance on high stakes tests

in third grade. *Educational Evaluation and Policy Analysis*, *32*(2), 166–182. https://doi.org/10.3102/0162373710367682

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263. https://doi.org/10.1111/rssb.12027

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving causal inference: Recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research*, *8*(2), 279–303. https://doi.org/10.1086/691464

Kelcey, B. M. (2009). *Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings* [Doctoral dissertation, University of Michigan]. https://deepblue.lib.umich.edu/handle/2027.42/63716.

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

Kim, J., & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection processes vary across schools*. Los Angeles, CA: Center for Study of Evaluation. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.113.4052

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). Guilford.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*(1), 63–82. https://doi.org/https://doi.org/10.3102/00346543075001063

Kupchik, A. (2010). *Homeroom security: School discipline in an age of fear*. NYU Press.

Kutsyuruba, B., Klinger, D. A., & Hussain, A. (2015). Relationships among school climate, school safety, and student achievement and well-being: A review of the literature. *Review of Education*, *3*(2), 103–135. https://doi.org/10.1002/rev3.3043

Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight Trimming and Propensity Score Weighting. *PLOS ONE*, *6*(3), e18174. https://doi.org/10.1371/journal.pone.0018174

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265–284. https://doi.org/10.1080/00273171.2014.991018

Leyrat, C., Caille, A., Donner, A., & Giraudeau, B. (2013). Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in Medicine, 32*(19), 3357–3372. https://doi.org/10.1002/sim.5795

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., … Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical Methods in Medical Research, 28*(1), 3–19. https://doi.org/10.1177/0962280217713032

Lingle, J. A. (2009). *Evaluating the performance of propensity scores to address selection bias in a multilevel context: A Monte Carlo simulation study and application using a national dataset* [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/eps_diss/56/

Lorah, J. (2018). Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(8). 1-11. https://doi.org/10.1186/s40536-018-0061-2

Lüdecke, D. (2019). sjstats: Statistical Functions for Regression Models [R package]. doi: 10.5281/zenodo.1284472

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. https://doi.org/10.1037/a0012869

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

May, D. C., Fessel, S. D., & Means, S. (2004). Predictors of principals' perceptions of school resource officer effectiveness in Kentucky. *American Journal of Criminal Justice*, *29*(1), 75-93. https://doi.org/10.1007/BF02885705

McCormick, M. P., O'Connor, E. E., Cappella, E., & McClowry, S. G. (2013). Teacher–child relationships and academic achievement: A multilevel propensity score model approach. *Journal of School Psychology, 51*(5), 611–624. https://doi.org/10.1016/j.jsp.2013.05.001

McDevitt, J., & Panniello, J. (2005). *National assessment of school resource officer programs: Survey of students in three large new SRO programs*. Washington, DC: U.S. Department of Justice. Retrieved September, 19, 2019 from https://files.eric.ed.gov/fulltext/ED486271.pdf.

Minneapolis Board of Education (2020, June 2). *Resolution to terminate the contract for services with the Minneapolis Police Department for the services of school resource officers* [Resolution 2020-0037]. https://v3.boardbook.org/Public/PublicItemDownload.aspx?ik=46459777

Minnesota Department of Education. (2019a). *Minnesota student survey*. Retrieved September 19, 2019 from http://education.state.mn.us/MDE/dse/health/mss.

Minnesota Department of Education. (2019b). *Data reports and analytics* [Dataset]. Retrieved September 19, 2019 from http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

Musu, L., Zhang, A., Wang, K., Zhang, J., & Oudekerk, B. (2019). *Indicators of school crime and safety: 2018.* (NCES 2019-047/NCJ 252571). National Center for Education Statistics, U.S. Department of Education, and Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice. Washington, DC. https://www.bjs.gov/content/pub/pdf/iscs18.pdf

National Association of School Resource Officers (2020). *General FAQs*. Retrieved May 28, 2020 from https://nasro.org/handle/20.500.11990/1242frequently-asked-questions/

Nickerson, A. B., & Martens, M. (2008). School violence: Associations with control, security/enforcement, educational/therapeutic approaches, and demographic factors. *School Psychology Review, 37*(2), 228-243. https://doi.org/10.1080/02796015.2008.12087897

Nickodem, K. Rodriguez, M. C., Lamm, R., & Park, K. (2019, April). *Social and emotional learning ICCs and associations with school composition and achievement* [Paper presentation]. National Council on Education in Measurement annual meeting, Toronto, Canada.

Noguera, P. A. (2003). Schools, prisons, and social implications of punishment: Rethinking disciplinary practices. *Theory into Practice, 42*(4), 341-350. https://doi.org/10.1207/s15430421tip4204_12

Office of Civil Rights. (2019). *Civil rights data collection* [Dataset]. Retrieved September 19, 2019 from https://ocrdata.ed.gov/

Pirracchio, R., Resche-Rigon, M., & Chevret, S. (2012). Evaluation of the Propensity score methods for estimating marginal odds ratios in case of small sample size.

BMC Medical Research Methodology, *12*(1), 70. https://doi.org/10.1186/1471-2288-12-70

Premo, J., Lamb, R. & Cavagnetto, A. (2018). Conditional cooperators: Student prosocial dispositions and their perceptions of the classroom social environment. *Learning Environments Research, 21*, 229–244. https://doi.org/10.1007/s10984-017-9251-z

R Core Team (2020). *R: A language and environment for statistical computing* [software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

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

Reardon, S. F., Kalogrides, D., & Shores, K. (2019). The Geography of Racial/Ethnic Test Score Gaps. *American Journal of Sociology*, *124*(4), 1164–1221. https://doi.org/10.1086/700678

Robers, S., Kemp, J., & Truman, J. (2013). *Indicators of school crime and safety: 2012.* (NCES 2013-036/NCJ 241446). National Center for Education Statistics, U.S. Department of Education, and Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice. Washington, DC. https://eric.ed.gov/?id=ED543705

Rodriguez, M. C. (2017). *2013-2016 Minnesota student survey technical report on measures of developmental skills, supports, & challenges*. University of Minnesota Digital Conservancy. http://hdl.handle.net/11299/195197.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.2307/2335942

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. https://doi.org/10.1080/01621459.1984.10478078

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

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3–4), 169–188. https://doi.org/10.1023/A:1020363010465

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, *2*(3), 808–840. https://doi.org/10.1214/08-AOAS187

Schunck, R. (2016). Cluster size and aggregated level 2 variables in multilevel models. A cautionary note. *methods, data, analyses*, *10*(1), 97-108. https://doi.org/10.12758/mda.2016.005

Servoss, T. J. (2017). School security and student misbehavior: A multi-level examination. *Youth & Society*, *49*(6), 755–778. https://doi.org/10.1177/0044118X14561007

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

Shaw, E. J., & Mattern, K. D. (2009). *Examining the accuracy of self-reported high school grade point average* (No. 2009–5). College Board. https://eric.ed.gov/?id=ED562616

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72*(1), 146-148. https://doi.org/10.1037/0021-9010.72.1.146

Snijders, T. & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and applied multilevel analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Snyder, T., & Musu-Gillette, L. (2015, April 16). Free or reduced price lunch: A proxy for poverty? [Blog post]. National Center for Education Statistics. https://nces.ed.gov/blogs/nces/post/free-or-reduced-price-lunch-a-proxy-for-poverty

Sticca, F., Goetz, T., Bieg, M., Hall, N. C., Eberle, F., & Haag, L. (2017). Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study. *PLoS ONE*, *12*(11). https://doi.org/10.1371/journal.pone.0187367

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

Tanner-Smith, E. E., & Fisher, B. W. (2016). Visible school security measures and student academic performance, attendance, and postsecondary aspirations. *Journal of Youth and Adolescence*, *45*(1), 195-210. http://dx.doi.org/10.1007/s10964-015-0265-5

Theriot, M. T., & Cuellar, M. J. (2016). School resource officers and students' rights. *Contemporary Justice Review, 19*(3), 363-379. http://dx.doi.org/10.1080/10282580.2016.1181978

Theriot, M. T., & Orme, J. G. (2016). School resource officers and students' feelings of safety at school. *Youth Violence and Juvenile Justice, 14*(2), 130-146. http://dx.doi.org/10.1177/1541204014564472

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

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

United States Department of Education. (2014). *Guiding principles: A resource guide for improving school climate and discipline*. http://www2.ed.gov/policy/gen/guid/school-discipline/guiding-principles.pdf

Van Boekel, M., Bulut, O., Stanke, L., Palma Zamora, J. R., Jang, Y., Kang, Y., & Nickodem, K. (2016). Effects of participation in school sports on academic and social functioning. *Journal of Applied Developmental Psychology, 46*, 31–40. https://doi.org/10.1016/j.appdev.2016.05.002

Voight, A., Hanson, T., O'Malley, M., Adekanye, L. (2015). The racial school climate gap: within-school disparities in students' experiences of safety, support, and connectedness. *American Journal of Community Psychology, 56*, 252–267. https://doi.org/10.1007/s10464-015-9751-x

Weiler, S. & Cray, M. (2011). Police at school: A brief history and current status of school resource officers. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 84*(4), 160-163. https://doi.org/10.1080/00098655.2011.564986

What Works Clearinghouse (2020). *Standards Handbook* (Version 4.1). Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, & Evaluation, 18*(11), 1-14. https://doi.org/10.7275/55hn-wk47

Wyse, A. E., Keesler, V., & Schneider, B. (2008). Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. *Teachers College Record, 110*(9), 1879-1900.

Yu, B. (2012). *Variable selection and adjustments in relation to propensity scores and prognostic scores: From single-level to multilevel data* [Doctoral dissertation, University of Toronto]. http://hdl.handle.net/1807/32855

Zakrison, T. L., Austin, P. C., & McCredie, V. A. (2018). A systematic review of propensity score methods in the acute care surgery literature: Avoiding the pitfalls and proposing a set of reporting guidelines. *European Journal of Trauma and Emergency Surgery*, *44*(3), 385–395. https://doi.org/10.1007/s00068-017-0786-6

Appendix A

Effect Size (Partial-ω^2) of Association Between Manipulated Factors (Rows) and Sample Characteristic (Columns)

	Correlation				Variance				ICC(1)		(2)
Factor	L1	Tr L2	Agg L2	Tr-Agg L2	Y	L1	Tr L2	Agg L2	Y	X	X
Subjects (A)	.00	**.86**	.09	**.50**	**.96**	.00	**.99**	**.85**	**.97**	.08	**1.00**
Clusters (B)	.00	.08	.00	.00	**.70**	**.96**	**.92**	**.31**	**.69**	**.80**	**.95**
ICC(1) (C)	.00	**.83**	.11	**.87**	**1.00**	**.94**	**1.00**	**.98**	**1.00**	**1.00**	**1.00**
Error Magnitude (D)	.00	-.01	**.83**	**.98**	.00	.00	.00	**1.00**	.00	.00	-.01
Procedure (E)	**1.00**	**1.00**	**.97**	**.97**	**1.00**	.00	**1.00**	**1.00**	**1.00**	.01	.00
A × B	.00	.01	-.01	-.01	.06	-.01	.04	-.01	.03	.09	**.85**
A × C	.01	**.44**	.00	.05	**.24**	.01	**.42**	.02	**.27**	.03	**1.00**
A × D	.00	-.01	.02	.10	.00	.01	-.01	**.41**	.00	.00	-.01
A × E	-.01	**.87**	**.84**	**.72**	**.96**	.00	**.99**	**.71**	**.97**	.00	-.01
B × C	.01	-.01	-.01	-.01	**.25**	**.88**	**.72**	.01	.09	**.71**	**.81**
B × D	.00	-.01	-.01	-.01	.00	.00	-.01	**.16**	-.01	.00	-.01
B × E	**.63**	.02	-.01	-.01	**.31**	.00	.07	.09	**.60**	.00	-.01
C × D	-.01	-.01	.08	**.44**	-.01	.00	-.01	.00	.00	.00	-.01
C × E	**1.00**	**.83**	**.83**	**.42**	**1.00**	.01	**.59**	.01	**1.00**	.01	-.01
D × E	-.01	-.02	**.92**	**.94**	.00	.00	-.02	**1.00**	.00	.01	-.02

Continued on Next Page

Continued

Factor	Propensity Score Estimation				%
	Convergence	Bias	MAE	RMSE	Treated
Subjects (A)	.00	**.57**	.01	.03	**.62**
Clusters (B)	**1.00**	.00	**.73**	**.89**	.00
ICC(1) (C)	**.42**	**.97**	**.76**	**.70**	**.98**
Error Magnitude (D)	**.45**	.01	**.74**	**.77**	.00
Procedure (E)	.32	.09	**.66**	**.69**	**.23**
A × B	.00	.00	.07	.08	.00
A × C	-.01	**.27**	.00	.00	**.22**
A × D	.01	.00	.04	.05	.00
A × E	.05	**.55**	**.17**	**.17**	**.61**
B × C	**.32**	-.01	.07	.10	-.01
B × D	**.44**	.00	**.50**	**.49**	.00
B × E	**.35**	.00	**.40**	**.38**	-.01
C × D	.04	-.01	**.14**	**.17**	-.01
C × E	.05	**.43**	**.21**	**.23**	**.57**
D × E	**.17**	.00	**.51**	**.53**	-.01

Note. Strict interpretation of the values is not advised because of violations of ANOVA assumptions. Values are intended to provide a rough indicator of the possible association with the dependent variable. L1 = 20 Level 1 Covariates; L2 = 10 Level 2 Covariates; Y = Outcome; X = 10 Level 1 Covariates aggregated to Level 2; Tr = True; Agg = Aggregated; ICC = Intraclass Correlation; MAE = Mean Absolute Error; RMSE = Root Mean Standard Error; **Bold indicates large effect size (> .14).**

Appendix B

Effect Size (Partial-ω^2) of Association Between Manipulated Factors (Rows) and Dependent Variables (Columns)

	Convergence Rate			Number of		Tr-Agg L2 Corr-	Agg ICC		Baseline Samples			
									ASD		VR	
Factor	PS	Y	Baseline	Subjects	Clusters	elation	(1)	(2)	Mean	Count	Mean	Count
Subjects (A)	.02	.00	**.33**	**.98**	.05	**.95**	**.54**	**1.00**	**.97**	**.99**	**.87**	**.26**
Clusters (B)	**.52**	.13	**.41**	**.96**	**.98**	.01	**.31**	**.53**	**.97**	**.99**	**1.00**	**.99**
ICC(1) (C)	**.15**	.01	.06	.03	.04	**1.00**	**1.00**	**1.00**	**.98**	**.99**	**.70**	**.66**
Error Magnitude(D)	**.48**	.03	**.54**	**.30**	**.44**	**1.00**	**.83**	**.67**	**.97**	**.48**	**.48**	**.42**
PS Model (E)	.00	.08	.00	.01	**.86**	.00	**.24**	.06	**.17**	**.23**	.00	.00
Conditioning (F)	.00	**.16**	.00	**.96**	**.93**	**.69**	**.93**	**.86**	.00	.00	.00	.00
A × B	.04	-.01	.06	**.87**	-.01	-.01	.00	**.24**	**.36**	**.80**	**.25**	.04
A × C	-.01	-.01	**.24**	.00	-.01	**.81**	.01	**.99**	**.80**	**.95**	.01	.14
A × D	.04	-.01	**.27**	.07	.01	**.75**	**.15**	**.44**	**.40**	.07	.05	.02
A × E	.00	.00	.00	.00	.03	.00	.02	.08	.01	.01	.00	.00
A × F	.00	.00	.00	**.89**	.02	.02	**.35**	**.72**	.00	.00	.00	.00
B × C	**.26**	.00	-.01	-.01	-.01	-.01	**.22**	.10	.02	**.41**	.11	**.22**
B × D	**.65**	.14	.08	.00	.01	-.01	**.15**	.10	**.68**	**.53**	**.51**	**.46**
B × E	.00	.09	.00	.00	**.39**	.00	.01	.01	.10	.11	.00	.00
B × F	.00	.06	.00	**.70**	**.48**	.00	.00	**.16**	.00	.00	.00	.00
C × D	**.23**	.00	.04	.05	.05	**.98**	**.55**	.04	**.58**	**.43**	**.27**	**.18**
C × E	.00	.00	.00	.00	.00	.00	.11	.00	.09	.10	.00	.00
C × F	.00	.01	.00	.00	.00	.00	**.79**	**.28**	.00	.00	.00	.00
D × E	-.01	.06	.01	-.01	.00	.00	**.16**	.03	.06	.13	.00	-.01
D × F	-.01	.04	-.01	.10	**.24**	**.32**	**.56**	**.51**	-.01	-.01	.00	.00
E × F	.00	.08	.00	.01	**.86**	.00	**.17**	.05	.00	.00	.00	.00

Continued on Next Page

	Propensity Score Samples							Baseline Samples		
	ASD		VR		Estimation			Estimation		
Factor	Mean	Count	Mean	Count	Bias	MAE	RMSE	Bias	MAE	RMSE
Subjects (A)	**.46**	**.52**	.00	.02	**.61**	**.32**	.00	**.93**	**.94**	**.95**
Clusters (B)	**.88**	**.89**	.01	**.98**	.00	**.24**	.11	.00	**.42**	**.78**
ICC(1) (C)	**.78**	**.81**	.00	.08	**.62**	**.44**	.02	**.94**	**.96**	**.97**
Error (D)	**.72**	**.68**	.00	**.86**	**.97**	**.82**	.00	**1.00**	**1.00**	**1.00**
PS Model (E)	**.95**	**.97**	.01	**.68**	.00	.09	.05	.00	.00	.00
Conditioning (F)	**.79**	**.78**	.01	.12	.00	.10	.05	.00	.00	.00
A × B	.00	.00	.00	-.01	-.01	-.01	-.01	-.01	-.01	.00
A × C	.01	.04	.01	.00	.09	.03	-.01	**.56**	**.58**	**.60**
A × D	.00	.00	.00	.05	**.54**	**.17**	.00	**.91**	**.91**	**.91**
A × E	.00	.02	.00	.00	.00	.00	.00	.00	.00	.00
A × F	.01	.02	.00	.00	.00	.00	.00	.00	.00	.00
B × C	.00	.00	.00	.00	.01	.02	.00	-.01	.03	.10
B × D	.00	.04	-.01	**.18**	-.01	**.16**	.07	-.01	**.33**	**.15**
B × E	**.48**	**.38**	.01	**.33**	.00	.13	.10	.00	.00	.00
B × F	.01	.01	.01	**.24**	.00	.13	.10	.00	.00	.00
C × D	**.16**	**.21**	.00	**.22**	**.57**	.10	-.01	**.92**	**.89**	**.88**
C × E	**.34**	**.34**	.00	.00	.01	.01	.00	.00	.00	.00
C × F	**.21**	**.22**	.00	.00	.01	.01	.00	.00	.00	.00
D × E	**.27**	**.47**	.00	.03	.00	.05	.03	-.01	-.01	-.01
D × F	**.55**	**.48**	.00	.08	.00	.06	.03	-.01	-.01	-.01
E × F	**.75**	**.84**	.01	**.66**	.00	.09	.05	.00	.00	.00

Note. Strict interpretation of the values is not advised because of violations of ANOVA assumptions. Values are intended to provide a rough indicator of the possible association with the dependent variable. Y = Outcome; Tr = True; Agg = Aggregated; ICC = Intraclass Correlation; PS = Propensity Score; MAE = Mean Absolute Error; RMSE = Root Mean Standard Error; **Bold indicates large effect size (> .14).**

Appendix C

Covariate Descriptive Statistics in Full Sample

	Mean (*SD*) or %		
	Total	SRO	No SRO
Student - Demographics			
8th Grade	37	34	49
9th Grade	35	36	28
11th Grade	29	30	23
Female	51	51	50
American Indian	1	1	1
Asian	6	6	3
Black	5	5	3
White	72	71	78
Multiracial	7	7	7
Latinx	9	9	9
Special Education	9	9	11
Free/Reduced Lunch	26	25	28
Student – Out-of-School Time			
Sports Team	59	58	62
School Club	27	27	27
Community Club	9	9	12
Tutoring	12	13	10
Leadership	13	13	13
Artistic Lessons	21	21	21
Physical Lessons	25	25	24
Religious Activity	37	36	43
Student – Mental & Physical Health			
Experienced Trauma	37	37	39
Bullied	6.99 (1.34)	6.96 (1.34)	7.11 (1.36)
Mental Distress	7.02 (1.37)	7.03 (1.37)	7.02 (1.36)
Positive Identity	11.13 (1.87)	11.14 (1.87)	11.09 (1.85)
Social Competence	11.38 (1.63)	11.4 (1.64)	11.24 (1.61)
Family/Community Support	12.29 (1.85)	12.3 (1.85)	12.27 (1.84)
College Aspiration	80	81	77
Safe Travel	97	97	97
Poor Health	1.11 (0.93)	1.10 (0.93)	1.12 (0.92)
Sick 3 Days	9	9	9

Continued on Next Page

	Mean (*SD*) or %		
	Total	SRO	No SRO
Skipped Meal	4	4	4
Eat Produce Daily	34	34	30
Sleep 8 Hours Daily	38	37	44
Student – Risky Behaviors (in last 30 days)			
Skipped Class	16	16	14
Gambled	32	32	32
Vandalized	12	12	13
Stole	9	9	8
Smoked Cigarette	5	4	6
Smoked eCig	10	10	8
Smoked Marijuana	8	9	6
Drank Alcohol	14	14	14
Took Rx	5	5	4
School - Characteristics			
Total Students	1137.9 (708.8)	1281.0 (692.4)	479.2 (283.1)
Includes 5th grade	5	4	9
Includes 8th grade	51	44	86
Includes 9th & 11th grade	72	72	72
Charter/Magnet	7	8	3
Urban	52	57	27
% Female	0.49 (0.02)	0.49 (0.02)	0.48 (0.03)
% American Indian	0.01 (0.06)	0.01 (0.06)	0.01 (0.04)
% Asian	0.06 (0.07)	0.06 (0.07)	0.03 (0.05)
% Latino	0.07 (0.08)	0.07 (0.07)	0.07 (0.1)
% Black	0.07 (0.09)	0.08 (0.09)	0.04 (0.08)
% White	0.76 (0.19)	0.74 (0.20)	0.83 (0.17)
% Multiracial	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)
% Free/Reduced Lunch	0.31 (0.17)	0.30 (0.17)	0.34 (0.15)
% Special Education	0.13 (0.03)	0.12 (0.03)	0.14 (0.04)
% Eng. Lang. Learner	0.04 (0.06)	0.04 (0.06)	0.03 (0.06)
School – Out-of-School Time			
% Sports Team	0.58 (0.09)	0.57 (0.08)	0.61 (0.09)
% School Club	0.27 (0.07)	0.27 (0.07)	0.27 (0.08)
% Community Club	0.10 (0.04)	0.09 (0.03)	0.12 (0.06)
% Tutoring	0.13 (0.06)	0.13 (0.06)	0.10 (0.06)
% Leadership	0.13 (0.04)	0.13 (0.04)	0.13 (0.05)
% Artistic Lessons	0.21 (0.05)	0.21 (0.05)	0.21 (0.07)
% Physical Lessons	0.25 (0.08)	0.25 (0.08)	0.24 (0.10)
% Religious Activity	0.36 (0.10)	0.35 (0.09)	0.41 (0.12)
Positive OST Experiences	11.16 (0.35)	11.18 (0.34)	11.07 (0.35)

Continued on Next Page

	Mean (*SD*) or %		
	Total	SRO	No SRO
School – Mental & Physical Health			
% Experienced Trauma	0.37 (0.09)	0.37 (0.08)	0.40 (0.09)
Bullied	7.01 (0.23)	6.99 (0.23)	7.14 (0.22)
Mental Distress	7.03 (0.18)	7.03 (0.16)	7.03 (0.26)
Positive Identity	11.10 (0.31)	11.11 (0.30)	11.06 (0.34)
Social Competence	11.35 (0.36)	11.37 (0.34)	11.22 (0.4)
Family/Community Support	12.27 (0.36)	12.27 (0.36)	12.26 (0.38)
% Safe Travel	0.96 (0.03)	0.96 (0.03)	0.96 (0.02)
Poor Health	1.12 (0.15)	1.12 (0.14)	1.13 (0.16)
% Sick 3 Days	0.09 (0.03)	0.09 (0.03)	0.09 (0.03)
% Skipped Meal	0.05 (0.02)	0.05 (0.02)	0.04 (0.03)
% Eat Produce Daily	0.33 (0.07)	0.34 (0.07)	0.30 (0.07)
% Sleep 8 Hours Daily	0.39 (0.12)	0.37 (0.11)	0.46 (0.11)
School – Risky Behaviors (in last 30 days)			
% Skipped Class	0.17 (0.08)	0.17 (0.08)	0.15 (0.07)
% Gambled	0.32 (0.05)	0.32 (0.05)	0.32 (0.07)
% Vandalized	0.12 (0.04)	0.12 (0.03)	0.13 (0.05)
% Stole	0.09 (0.04)	0.09 (0.04)	0.08 (0.05)
% Smoked Cigarette	0.05 (0.04)	0.05 (0.03)	0.06 (0.05)
% Smoked eCig	0.10 (0.06)	0.10 (0.06)	0.08 (0.06)
% Smoked Marijuana	0.08 (0.05)	0.09 (0.05)	0.06 (0.04)
% Drank Alcohol	0.14 (0.07)	0.14 (0.06)	0.14 (0.07)
% Took Rx	0.05 (0.02)	0.05 (0.02)	0.04 (0.03)
School - Academics			
% College Aspiration	0.78 (0.07)	0.79 (0.07)	0.75 (0.08)
% Proficient - Math	0.55 (0.16)	0.56 (0.15)	0.53 (0.16)
% Proficient - Reading	0.61 (0.12)	0.62 (0.12)	0.58 (0.12)
School - Resources			
Teacher Absences	0.33 (0.20)	0.34 (0.19)	0.27 (0.21)
Student/Teacher Ratio	17.06 (3.43)	17.56 (3.30)	14.74 (3.03)
Security Guard	19	23	4
Counselor	95	97	86
Nurse	89	93	74
Psychologist	73	77	56