10-16-2009

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

Jeremy Andrew Lingle
Georgia State University

ACCEPTANCE

This dissertation, 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, by JEREMY ANDREW LINGLE, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Phill Gagné, Ph.D.
Committee Chair

_____
Carolyn Furlow, Ph.D.
Committee Member

_____
Philo Hutcheson, Ph.D.
Committee Member

_____
Sheryl A. Gowen, Ph.D.
Committee Member

_____
Deanne Swan, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
R. W. Kamphaus, Ph.D.
Dean and Distinguished Research Professor
College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Jeremy Andrew Lingle

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Jeremy Andrew Lingle
3522 Beachhill Drive
Atlanta, GA 30340

The director of this dissertation is:

Dr. Phill Gagné
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303 – 3083

VITA

Jeremy Andrew Lingle

ADDRESS: 3522 Beachill Drive
Atlanta, Georgia 30340

EDUCATION:

Ph.D.	2009	Georgia State University Educational Policy Studies
M.S.	2005	Georgia State University Counseling and Psychological Services
B.S.	1998	University of Alabama Psychology

PROFESSIONAL EXPERIENCE:

2009–present	Project Director EMSTAR Research, Inc., Atlanta, GA
2006–2009	Research Coordinator Georgia State University, Atlanta, GA
2003–2006	Graduate Research Assistant Georgia State University, Atlanta, GA

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2005–present	American Evaluation Association
2005–present	American Educational Research Association

PRESENTATIONS AND PUBLICATIONS:

Lingle, J., Alemdar, M., & Gowen, S. (April 2009). Comparing Methods of Propensity Score Estimation with Partially Missing Data. Paper presented at the Annual Conference for the American Educational Research Association, San Diego, CA.

Lingle, J., & Gowen, S. (October 2008). Linking Academic Achievement to Attendance in Georgia's 21st Century Community Learning Centers Programs: A Propensity Score Demonstration. Paper presented at the annual National Evaluation Institute hosted by CREATE. Wilmington, NC.

Lingle, J., Skelton, S., and Davis, C. (April, 2008). Findings from the Behavior and Risk in Teens Survey. Presentation at the State Meeting of Teen Health Center Coordinators hosted by Georgia DHR.

Lingle, J., Alemdar, M., Gowen, S., & Skelton, S. (March 2008). School Engagement, After School Activities, and Health-Risk Behaviors: Results from an Evaluation of Community-Based After School Programs. Paper presented at the Annual Conference for the American Educational Research Association, New York, NY.

Suprina, J. & Lingle, J. (2008). Overcoming societal discouragement: Recovering alcoholics' perceptions of the Adlerian life tasks. *Journal of Individual Psychology, 64*(2). 193-212.

Lingle, J., Furlow, C., Gowen, S., & Skelton, S. (November 2007). The Application of Multi-Level Modeling in the Evaluation of After-school Programs: Linking Academic Success to Attendance. Paper presentation at the Annual Conference for the American Evaluation Association, Baltimore, MD.

ABSTRACT

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

by

JEREMY A. LINGLE

When researchers are unable to randomly assign students to treatment conditions, selection bias is introduced into the estimates of treatment effects. Random assignment to treatment conditions, which has historically been the scientific benchmark for causal inference, is often impossible or unethical to implement in educational systems. For example, researchers cannot deny services to those who stand to gain from participation in an academic program. Additionally, students select into a particular treatment group through processes that are impossible to control, such as those that result in a child dropping-out of high school or attending a resource-starved school. Propensity score methods provide valuable tools for removing the selection bias from quasi-experimental research designs and observational studies through modeling the treatment assignment mechanism.

The utility of propensity scores has been validated for the purposes of removing selection bias when the observations are assumed to be independent; however, the ability of propensity scores to remove selection bias in a multilevel context, in which group membership plays a role in the treatment assignment, is relatively unknown. A central purpose of the current study was to begin filling in the gaps in knowledge regarding the performance of propensity scores for removing selection bias, as defined by covariate balance, in multilevel settings using a Monte Carlo simulation study. The performance of propensity scores was also examined using a large-scale national dataset.

Results from this study provide support for the conclusion that multilevel characteristics of a sample have a bearing upon the performance of propensity scores to balance covariates between treatment and control groups. Findings suggest that propensity score estimation models should take into account the cluster-level effects when working with multilevel data; however, the numbers of treatment and control group individuals within each cluster must be sufficient to allow estimation of those effects. Propensity scores that take into account the cluster-level effects can have the added benefit of balancing covariates within each cluster and across the sample as a whole.

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

by

Jeremy A. Lingle

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Research, Measurement, and Statistics
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2009

ACKNOWLEDGEMENTS

This dissertation was accomplished through the guidance and support of many people over many years. I begin by expressing my gratitude to my dissertation committee chair, Dr. Phill Gagné, who sacrificed much time and energy to the crafting of this dissertation. Thank you for keeping me on track, for providing guidance and encouragement, and for challenging my interpretations. This study would not have been possible without your contribution. I would like to thank the chair of the EPS department, Dr. Sheryl Gowen, who not only agreed to be a member of my committee, but diligently fought for funding for me during my academic career, and has always offered her support as my journey continues. The resulting evaluation and research experiences that I gained under her tutelage are invaluable. I am also extremely grateful for Dr. Carolyn Furlow, who served as my doctoral chair throughout my coursework, who had great influence on my research interests which culminated in this dissertation and choice of career, and who served on my dissertation committee. I would like to thank Dr. Deanne Swan for serving on my dissertation committee, helping me explore the many details of propensity score adjustment methods and taking the time to listen and guide me as I navigated numerous dissertation dilemmas. Dr. Philo Hutcheson also deserves my gratitude for his contributions on my committee and helping to make this study more valuable for educational researchers. He also deserves an award for (yet again) venturing into the strange world of simulation studies and quantitative research. I would also like to acknowledge Dr. Bill Curlette for his contribution to my journey into this doctoral program. His example and explanations not only helped to demystify statistics for me, but also helped me to understand how to meld my interests in counseling and research. I am grateful that he guided me to this program of study. My parents, Dot and Don Lingle, unquestionably deserve much of the credit for the successful completion of this dissertation. Because of their love and support, I had the confidence to face and accomplish the many challenges that come with educational pursuits. They also provided me with the love of learning that continues to give me so much happiness (I really am "Dr. J. Andrew" now, Dad!). Dr. Meltem Alemdar: my guide, my graduate-school compatriot, my friend. Thank you for listening to my frustrations, warning me of approaching hurdles, lifting me up at my weak times, and always keeping me grounded. Probably most importantly, though, I thank you for being the honest, loyal, genuine, and caring person that you are. Finally, I want to thank my partner, Jeremy Brasseal, who I promised long ago that, one day, I would be through with school. Thank you for your patience, support, and love along this long and winding road.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

PROPENSITY SCORES IN AN EDUCATIONAL CONTEXT

When researchers are unable to assign students randomly to treatment conditions, selection bias is introduced into the estimates of treatment effects. Random assignment to treatment conditions, which has historically been the scientific benchmark for causal inference, is often impossible or unethical to implement in educational systems. For example, researchers cannot deny services to those who stand to gain from participation in an academic program. Additionally, students select into a particular treatment group through processes impossible to control, such as those that result in a child dropping out of high school or attending a resource-starved school. When a researcher is unable to assign an individual to a treatment condition, the factors that influence that student's exposure to a treatment must be modeled in order to estimate unbiased treatment effects. Propensity score methods provide valuable tools for removing the selection bias from quasi-experimental research designs.

The utility of propensity scores has been validated for the purposes of removing selection bias when the observations are assumed to be independent (Rosenbaum & Rubin, 1983); however, the ability of propensity scores to remove bias in a multilevel context, in which group membership plays a role in the decision to participate in a treatment, is relatively unknown. This study will explore the performance of propensity scores for removing selection bias, defined as balance on covariates between treatment groups, in a multilevel context.

This chapter begins with a discussion of causal inference using potential outcomes through Rubin's Causal Model (Holland, 1986). A description of propensity

scores and their estimation and application in a single-level context follows. The chapter concludes with a presentation of the challenges that multilevel data pose to the efficacy of propensity score methods for successfully addressing selection bias and enumerates the purposes of the current study in understanding those challenges.

Rubin's Causal Model

Through Rubin's Causal Model (Holland, 1986), causal effects of a treatment can be determined through comparison of the potential outcomes that would have been observed in an individual under different treatments. In its simplest form, an experimental treatment has two levels, one in which the treatment is administered to an individual, $Z = 1$, and one in which the treatment is not given to the individual, $Z = 0$. Such a formulation describes a basic treatment and control experimental design. In this scenario, two potential outcomes exist for an individual corresponding to the two levels of treatment. For a given experimental individual, *i*, these two potential outcome are designated as $Y_{i(Z=1)}$ for the outcome associated with the treatment condition and $Y_{i(Z=0)}$ for the outcome associated with the control condition (or simply $Y_{i(1)}$ and $Y_{i(0)}$, respectively). The treatment effect for an individual, Δ_i, is defined as the difference between these two outcomes, as described in Equation 1.1:

$$\Delta_i = Y_{i(1)} - Y_{i(0)} \tag{1.1}$$

The fundamental problem with this formulation is that only one outcome can be observed for an individual per treatment administration. Specifically, when treatment is administered to individual *i*, the observed outcome is $Y_{i(1)}$ and the missing outcome is $Y_{i(0)}$. In this scenario, $Y_{i(0)}$ is considered the counterfactual. The same logic applies when $Z = 0$, so that $Y_{i(0)}$ is observed and $Y_{i(1)}$ is the counterfactual. A researcher can never observe

both outcomes for any one individual simultaneously. The treatment effect, Δ_i, must be estimated from mean sample outcomes for $Z = 1$ and $Z = 0$, given by $Y_{(1)}$ and $Y_{(0)}$. This average treatment effect is given by:

$$\bar{\Delta} = \bar{Y}_{(1)} - \bar{Y}_{(0)}. \quad (1.2)$$

Treatment Assignment Mechanism. Rubin shows that this definition holds true when the assignment mechanism is known. This knowledge of how individuals are assigned to treatments is essential for causal inference. Comparing observed outcomes of those who received treatment and those who do not receive treatment is an unbiased estimate of the treatment effect only under a random treatment assignment mechanism. If the treatment assignment is not randomized, then the factors that affected assignment and are subsequently related to the reason why certain outcomes are missing must be included in the inferential model (Rubin, 1976).

In an experimental design that incorporates simple random sampling (SRS), all individuals have the same probability of receiving treatment. The researcher controls the treatment assignment mechanism, in that the researcher uses chance for assigning individuals to treatment or control groups. Importantly, the characteristics of the individual are not used in this assignment. In this instance, the sampling mechanism is known: Each individual in the sample has a 0.5 probability of receiving the treatment. This probability of receiving treatment can also be referred to as a propensity score. In the case of SRS, the true propensity score for each individual is equal to 0.5 (with the caveat that the propensity scores for individuals will vary from 0.5 as a result of sampling error). Because the assignment mechanism is based upon chance, the characteristics of the individuals in both the treatment and control group will be reflective of the sample

from which the individuals are drawn. Those characteristics, or covariates, will then be balanced between these groups. This relationship can be shown in the following expression, given the probability of receiving treatment assignment $p(Z = 1)$ and probability of control assignment $p(Z = 0)$ with a vector of pretreatment covariates, *x*:

$$p(Z = 1|x) = p(Z = 0|x) = 0.5. \quad (1.3)$$

In observational studies, the treatment assignment mechanism is often confounded by selection bias. Observational studies are those in which the treatment assignment mechanism is outside of the control of the researcher. In education, a commonly occurring observational study is one in which individuals self-select to participate in a program. Individuals who self-select to participate are likely different on important characteristics from individuals who select not to participate. Similarly, specific populations that are targeted for an intervention are systematically different from the general population. In such scenarios, the researcher does not control the factors influencing the receipt of treatment. The primary challenge then in observational studies is to make sure that those in the treatment and control groups are systematically dissimilar based upon characteristics that are related to participation in the treatment. If those characteristics that are imbalanced also have an effect upon the outcome, the estimation of a causal relationship between the treatment and the outcome will be confounded. For causal effect estimations, the influence of those factors upon treatment must be modeled and this can be accomplished through estimation of the individual's propensity score.

Addressing Challenges of Non-Randomized Experiments. One method of addressing imbalances between treatment and control groups is by comparing only those

individuals with like characteristics through matching or stratification. For example, if a researcher finds that the proportion of females and males is balanced in the population but is imbalanced between treatment and control groups in the sample, sex likely has an influence upon the treatment assignment mechanism and any estimation of average treatment effects will be influenced by selection bias. One simple solution to address such an imbalance is to divide the sample by the imbalanced characteristics (Rubin, 2008), in this case, by sex. Females in the treatment condition are compared to females in the control condition; likewise, males are compared to males.

This solution, often referred to as blocking, does not pose a problem if the sample size is large enough to accommodate such divisions. One challenge becomes obvious, however, if numerous covariates show evidence of imbalance between treatment and control groups. For example, in a case with imbalance in four dichotomous covariates, the number of divisions grows exponentially, with the required number of cells equaling 2^4, or requiring 16 comparisons of treatment and control groups. In addition to the problems associated with decreased sample size is the likelihood that members of either the treatment or control group might be absent from cells, preventing causal estimations among those with that particular combination of characteristics.

Propensity Scores

Rosenbaum and Rubin (1983, 1984) presented a method by which balance can be attained across multiple observed characteristics, given by *x*, through a single estimated propensity score. This propensity score is defined as the conditional probability of receiving treatment based upon *x*. Specifically, *x* represents a vector of measured pretreatment covariates that describe the individual that are related to the individual's

treatment assignment as well as to the treatment outcome. Using Rosenbaum and Rubin's formulation, this relationship is denoted by Equation 1.4 where *x* represents the individual's estimated propensity score:

$$x = p(z = 1|x) \tag{1.4}$$

Researchers can apply these propensity scores through methods of matching, stratification, covariate adjustment, or weighting in order to balance treatment and control groups on measured pretreatment covariates. Those individuals who have equal propensity scores but who are in different treatment conditions are, therefore, comparable.

Propensity Score Estimation. The estimation of propensity scores to account for selection bias is a multistep process that requires both theoretically and statistically based decision-making. Suggestions and advice for appropriate propensity score estimations are readily available in the research literature. Generally, the estimation process is as follows: (a) Choose a statistical method to estimate propensity scores; (b) Choose covariates to include in the estimation; and (c) Determine the balance between treatment and control group and adjust the model accordingly. The details and variations of these steps are further discussed in the following sections.

Choosing the Method for Estimation. The most commonly used method for estimating propensity scores is logistic regression. The vast majority of research using propensity scores focuses upon treatments that are dichotomous, typically with individuals either as a member of a treatment or control group. Propensity score methods are also applicable for the comparison of multiple treatment groups and have been

expanded to address ordinal treatment exposures, such as the case with levels of dosage (Imbens, 2000; Joffe & Rosenbaum, 1999; Lu et al., 2001).

The form of a logistic regression with a dichotomous outcome is given by Equation 1.5. In the case of treatment and control ($Z = \{0, 1\}$), the model can be used to estimate the probability, p_i, of an individual *i* receiving treatment conditional on his/her pretreatment covariates, *x*. Likewise, the probability that the individual is in the control group is $(1 - p_i)$. Logit or probit models are often used in order to eliminate predicted probabilities that might fall outside of these bounds and to provide a continuous and a more normally distributed outcome measure.

$$logit[p_i] = log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad (1.5)$$

In this equation, β_0 represents the average propensity score for the sample, or the average probability in logit units of receiving treatment across individuals; $\beta_{1\ldots n}$ represent the parameter estimates for individual-level predictors $x_{1\ldots n}$, or the influence that each of the individuals' characteristics have upon the probability of receiving treatment.

Selection of covariates. The literature surrounding propensity score methodology consistently offers two specific requirements for the choice of covariates to be included in the propensity score estimation: First, selected variables should be related to the treatment assignment mechanism and to the outcome measure; second, neither the treatment assignment nor the outcome should influence the predictor variables. For example, a student's ethnicity is not going to be influenced by participation in an academic program; however, a measure of self-esteem might be influenced by participation in the program. A measure of self-esteem, therefore, should be attained before the decision to participate in the program is made. This latter requirement has led

to these variables being commonly referred to as pretreatment covariates. Although this perspective is generally accepted, differing opinions are related to the roles that theory and statistics play in the decision for covariate selection, as well as the inclusion criteria for covariates that have inconclusive relationships with the treatment assignment and/or the outcome.

To assist in this discussion, let us consider a researcher's question: "Does access to a computer in the home have a positive effect upon academic achievement in mathematics?" In this study, the treatment assignment would be indicated by the student's access to computers in the home, either having that access (Z=1) or not having that access (Z=0). The outcome, Y, would be some continuous academic measure. In designing the study, the researcher would decide the best means by which to gather these data. If the researcher was unable to randomly assign students to a treatment or control condition, or such control of assignment was undesirable, the students would select into the treatment or control condition as a function of numerous individual and contextual factors. For example, the student's socioeconomic status or experience with technology may play a role in whether a computer is present in the home.

In many cases, the researcher must rely upon an existing environment from which to gather data, such as a school or classroom, or rely upon data that were previously collected based upon different priorities. For example, the Educational Longitudinal Study of 2002 (ELS:2002; NCES, 2004) is a large-scale dataset that contains many variables gathered from high school students, along with numerous contextual covariates describing their schools and their faculties. When collecting data from an existing

context, or when using data from an existing dataset, the researcher does not have control of who receives any given treatment.

When a researcher must collect data from participants rather than use an existing dataset, the choice of variables to collect must be made explicit prior to data collection. One must select those pretreatment covariates that are expected to affect both the treatment assignment and the outcome measure so that bias in the estimates of the treatment effect can be addressed. These decisions are based upon previous research in the particular field of study.

When data have been previously collected, however, as with the ELS:2002 dataset, and a finite number of variables are available, a researcher may be tempted to disregard theory and test the relationships between treatment assignment and each covariate through purely statistical means. A number of statistical procedures are readily available that will yield a statistically defendable list of covariates related to treatment assignment, such as stepwise regression. Likewise, the relationship of the covariates to the outcome measure can also be tested.

Rubin (2008), however, argues that reliance purely upon regression programs is "entirely inadequate" (p. 816), such as through the application of stepwise or backward selection techniques. Such techniques often result in insufficiently explicated theoretical models and unstated assumptions about covariate relationships with the assignment mechanism. Such practices do not test any particular scientific theory and are inadequate for inferences of causal effects.

Rubin (2008) offers a number of suggestions for designing successful experiments in an observational context: The researcher should think of the dataset that is to be used

in the observational study as having arisen from a hypothetical randomized experiment in which the assignment mechanism to the treatment condition has been lost. Before considering the actual covariates in the dataset, one should carefully consider why some individuals received the treatment and some did not. For the previous example, a researcher should consider the factors that might influence a family's computer ownership. Influential individual experiences and personality characteristics as well as community factors should be clarified. Rubin argues that it is this theoretically-based foundation that allows observational studies to be qualitatively comparable to classical randomized experiments.

Once key covariates are determined, a researcher should turn to the data, not to determine if those covariates are indeed related to treatment assignment, but to see if they have even been measured, and whether they have been measured well. Highly correlated proxy covariates may be included as surrogate predictors of treatment assignment for those covariates that were not measured. Rubin (2008) makes his perspective quite clear in stating "no amount of fancy analysis can salvage an inadequate data base" (p. 817). Finally, he suggests that statistical relationships with outcome measures should not be considered at all in the decision to include them in the estimation of propensity scores. The reasoning behind his argument is tied to the attempt to emulate randomized experiments with observational data. Just as a researcher conducting a randomized experiment would not have access to this outcome data, he suggests that observational researchers remove the outcome data from the data set during the design phase of the observational study.

The role of statistical decision-making is central in the selection of covariates throughout the propensity-based research literature. In estimation of the initial propensity score, this role is typically related to statistically significant relationships with the treatment assignment or to maximizing the prediction rate of the model. Often, researchers consider the joint information that a covariate is both significantly related to treatment assignment as well as contributing significantly to the prediction rates when making inclusion decisions (Heckman et al., 1998). When the rationale for covariate inclusion is based upon prediction, the goal is to acquire the best possible prediction of group membership. It is based upon the belief that unbiased estimates of treatment effects can be attained "if the selection model is fully known and perfectly measured" (Shadish, Luellen & Clark, 2006, p. 148).

The challenge of this rationale, however, is related to the necessity of propensity score overlap for causal inference as presented in Rubin's Causal Model. For a control individual to be comparable to a treated individual on an outcome, they must have a common propensity score. By sharing a common propensity score, the pretreatment covariates are balanced and the related selection bias is removed. If the participation in a program is perfectly predicted by the selected covariates, the propensity scores for each treatment group individual would be equal to one and for each control group members would be equal to zero. With a perfect model, no individuals would be comparable because no treatment and control individuals would have the same propensity score. For example, if a researcher found that use of computers in the classroom was perfectly predicted by socioeconomic status (SES), then those who were classified as lower SES would have no overlap on their propensity score with those who were classified as higher

SES. Including a measure for SES as a predictor of treatment assignment would account for all of the variability. Because of this consequence, some randomness is necessary (Heckman, Ichimura, & Todd, 1998). Studies that include covariates based purely on a prediction rationale have the potential to over-fit the treatment assignment model and eliminate the ability to find adequate comparisons.

Researchers who use statistical significance as a decision-making factor often incorporate iterative processes or consider relationships with treatment assignment bivariately. Rosenbaum (2002) argued that less dependence should be placed upon traditional statistical significance in the decision to include covariates. In order to capture covariates that might otherwise be excluded using traditional tests of significance, he suggesting including pretreatment covariates that show group differences of $|t| > 1.5$.

A challenge occurs when researchers use both theoretical and statistical methods for the construction of propensity scores with disagreeing results. In this case theory suggests that a predictor is important to treatment assignment whereas analyses show that the covariate is not statistically related. Studies by Heckman et al. (1997) and Dehejia and Wahba (1999) provide an argument that omission of important variables in the propensity score estimation can increase the bias whereas inclusion of extraneous variables can increase the variance and "exacerbates the support problem" (Caliendo & Kopeinig, 2008, p. 38). Rubin and Thomas (1996) suggest to err on the side of too many variables, suggesting that only variables that are believed to be unrelated to the outcome should be omitted.

A final rationale for selection of covariates is based upon the success of the propensity score in its ability to create balance between the treatment groups on key

covariates. Rosenbaum and Rubin (1984) offer the advice that any propensity score that serves to balance covariates across groups is sufficient, whether or not the model predicts the treatment group assignment with perfection. Tests for determining the quality of balance are related, not only to the model used to estimate the propensity score, but also the methods used to apply the propensity score to the observational data set.

Matching and Stratification on the Propensity Score

Matching. The attainment of balance of pretreatment covariates between treatment and control groups can be attained through propensity score matching or stratification techniques. The success of the propensity score model can be determined by the amount of bias that is removed as compared to the initial model or when compared to a predetermined limit of acceptable imbalance. If balance is not sufficiently achieved according to these standards, then the propensity score model is adjusted by including previously omitted covariates, interaction terms, or quadratic terms.

Gu and Rosenbaum (1993) argue that the choice of matching technique should be based on certain sample characteristics that affect the ease of matching. Samples in which a relatively large number of individuals are available in the pool of controls to be matched with a smaller number of treatment group members are considered to provide easier matches, whereas samples with relatively equal numbers of treatment and control individuals provide harder matching conditions. In their simulation study exploring different matching techniques for multiple covariates, Gu and Rosenbaum defined easy matches as having a 6:1 ratio of control group members to treatment group members. They also tested matching conditions that they considered to be moderate (4:1, 2:1) and difficult (1:1). A second sample characteristic that contributes to this ease of matching is

the separation between the propensity scores of the treatment and control group members, or, alternatively, the amount of overlap on the propensity scores. The amount of separation is an indicator of the bias in the selection mechanism. Samples with greater selection bias on key pretreatment covariates should show evidence of greater propensity score separation between the groups, therefore having smaller overlap on the propensity scores between the treatment and control groups and less available matches.

Simply defined, a matching algorithm is a method by which control individuals are assigned to treatment individuals based upon their distance on multivariate *X*. The choice of matching algorithm should be influenced by these previously stated sample characteristics. Gu and Rosenbaum evaluated the performance of two matching algorithms to determine their ability to match given different difficulty-levels of matching. The algorithms considered were greedy and optimal. Greedy matching algorithms choose to match pairs based on a "nearest neighbor" principle. Individuals in the treatment group are randomly ordered and matched with the first acceptable match from the pool of controls as defined by distance on the propensity scores. Optimal matching algorithms, however, consider the average distance between all matched pairs for the sample before selecting each subsequent match.

Finally, Gu and Rosenbaum (1993) considered the structure of the matching. The structures considered were 1:1 matching, 1:*k* matching, and full matching. These structure designations refer to the number of individuals that may be matched with another individual. For the 1:1 matching structure, each treatment group member is allowed to match with only one control individual; likewise only one control individual is allowed to match with a treatment individual. The 1:*k* matching structure allows multiple

control individuals to be matched to a single treated individual. The number of controls is a designated constant, *k*. Finally, full matching allows multiple matches for both treatment and control group members.

Gu and Rosenbaum (1993) offered suggestions based upon the findings from their simulation to inform researchers who use propensity score matching. Specifically, they found that under most conditions, both greedy and optimal matching successfully balanced propensity scores between treatment and control groups as well as resulted in small distances between individual pairs. Full matching outperformed other matching structures for bias reduction and balance attainment. All methods struggled in cases when the ratio of control individuals to treatment individuals was equal and when bias was large.

A final consideration for using propensity score matching techniques is in the definition of a tolerable range around a treated individual's propensity score within which acceptable matches with a control may be made. This predetermined range is referred to as a caliper. Cochran and Rubin (1973) provided evidence that matching within a 0.2 standard deviation caliper around a covariate results in the removal of 99% of the bias associated with that covariate; likewise, a 0.6 standard deviation caliper setting results in removal of 90% of the bias. The choice of caliper size, therefore, is directly related to the trade-off in the number of quality matches treatment-control matches and the number of treated individuals who will have no acceptable match among the control group members. No clear guidance on this choice of caliper width has been offered for matching on the propensity score (Oakes & Johnson, 2006). Applied studies and simulations using propensity score matching tend to use caliper values of between 0.01

and 0.05 units of the propensity score, itself. The width of this value should be based upon the confidence of the researcher in the predicted propensity scores.

Determining Balance in Matched Samples. The evaluation of a propensity score matching algorithm can be determined through either a measure of distance between matched pairs on multivariate *X* or through the balance achieved on covariates between the treatment groups. A small distance between matched pairs on *X* necessitates balance, but balance does not necessitate a small distance on matched pairs. This difference may be evidenced in examinations of the comparability of matched pairs versus the comparability of the groups as a whole. Although attaining a small distance between each matched pair is ideal, the ability to accomplish this task given a high-dimensional *X* may be impossible; however, a small distance on *X* may not be necessary for treatment and control groups to be comparable. Overall balance between the groups on *X*, either as a whole or in part, may be possible when a small distance is unattainable, and can allow unbiased estimations of treatment effects (Gu & Rosenbaum, 1993).

When matching on propensity scores, the initial measure of success is based upon the distances between propensity scores. The balance in the sample attained on the specific covariates from which the propensity score was estimated is then evaluated. The reduction in bias in pretreatment covariates can be determined using the standardized mean difference (SMD) (D'Agostino & Rubin, 2000; Rosenbaum & Rubin, 1985), given by Equation 1.6.

$$SMD = 100(\bar{X}_T - \bar{X}_C)/s_p \quad (1.6)$$

Here, $\bar{X}_T - \bar{X}_C$ is the difference between the mean value of a covariate for the control group subtracted from the mean value of that covariate for the treatment group;

and s_p is the pooled standard deviation of the covariate for the two groups, calculated as described in Equation 1.7. In this equation, $s_T^2 + s_C^2$ is the sum of the variances of the treatment group covariate and the control group covariate,

$$s_p = \sqrt{(s_T^2 + s_C^2)/2}. \quad (1.7)$$

Standardized mean difference values for a given covariate that surpass 10 have been used to indicate that imbalance remains on that covariate between the treatment and control groups (D'Agostino & Rubin, 2000; Rosenbaum & Rubin, 1985).

Stratification on the Propensity Score. Another commonly used method to attain balance between treatment and control groups is through stratification on the propensity score. Stratification of a sample into quintiles (at a minimum) based upon a continuous covariate has been shown to remove 90% of the bias associated with that covariate (Cochran, 1968). Likewise, Rosenbaum and Rubin (1983) showed that 90% of the bias in the variables used to estimate the propensity score can be removed through stratification on the propensity score. Stratification can be considered a coarse form of matching in which individuals are grouped according to their propensity scores. When stratifying on the propensity score, the range of scores within a particular stratum will be smaller than across the sample as a whole, resulting in treatment and control individuals with similar pretreatment covariates within each stratum. The goal of this procedure is to accomplish balance on the pretreatment covariates between the treatment and control groups within each stratum.

One benefit of stratification over matching is that it allows treatment individuals who might not have a close enough match among the control individuals on their propensity scores to be maintained in the sample as a member of the stratum. Methods of

stratification, however, require explicit consideration of the overlap of the propensity scores, whereas matching techniques do not. This area of propensity score overlap can be considered the region of common support where treatment and control individuals share common covariates so that unbiased effects can be estimated.

With matching, individuals who do not have acceptable matches in the comparison group are not included in the final matched sample. Such conditions often occur with treatment individuals who are in the positive tail of the propensity score distribution. These individuals have the greatest likelihood of receiving treatment but often have no comparable control individuals with equally high likelihood of receiving treatment. With stratification, the range of the propensity scores within the treatment and the control groups must be considered prior to stratifying to determine if the two groups have common support. If the distribution of the propensity score is not considered prior to stratification, the highest strata might consist of only treatment individuals and the lowest strata consist of only control individual, thereby preventing comparison in these strata.

The steps of stratification (also referred to as interval matching, blocking, and subclassification) are as follows: Once the initial propensity score for all individuals has been estimated, the region of overlap on the propensity scores between the treatment and control individuals is assessed. Those treated individuals and control individuals that have no comparisons are dropped from the sample. If sufficient overlap exists, the remaining sample is stratified into quintiles on the propensity score, and the balance on the propensity scores between the treatment and control groups is assessed within each strata. If imbalance remains, strata can be further divided and the balance reassessed.

Once balance is attained within each stratum, the balance on key pretreatment covariates is assessed within stratum. If imbalance on the covariates remains, the propensity score model may be augmented by including previously omitted covariates, interaction effects, and/or quadratic terms.

Determining Balance with Stratification on the Propensity Score. The determination of balance among the pretreatment covariates can be accomplished through two-way Analyses of Variance, with the treatment/control group assignment as one factor, the propensity score strata as the second factor, and a covariate as the dependent variable. In addition, the interaction between the treatment groups and the propensity score strata should be considered in determining the balance attained by a model. A positive interaction value suggests that the difference in the mean covariate value between treatment and control groups is larger among those individuals who are more likely to receive treatment. Such interactions are typically remedied through inclusion of interaction terms and/or non-linear functions of the covariates in the propensity score estimation model. If such imbalances or interactions cannot be remedied across all covariates, those confounders considered of greatest importance should show balance. Other strategies to address imbalanced covariates if other methods are unsuccessful is to divide the sample by the imbalanced covariate and conduct the analyses separately per group, or to remove the covariate from the propensity score estimation but include it as a blocking variable in the effect estimation.

Simulations Addressing Variable Selection for Propensity Score Estimation

A number of simulation studies have been conducted in order to provide guidance regarding the selection of pretreatment covariates to include in the propensity score

estimation. In a pair of simulation studies (Austin, 2008; Austin, Grootendorst, & Anderson, 2007), researchers evaluated the performance of a number of propensity score models. In each of these studies, the simulated relationships between predictors, treatment assignment, and the outcome measure can be described through the following table.

Table 1:

Relationships of Covariates to Treatment Assignment and Outcome

Outcome	Treatment		
	Strongly Associated	Moderately Associated	Not Associated
Strongly Associated	x_1	x_2	x_3
Moderately Associated	x_4	x_5	x_6
Not Associated	x_7	x_8	x_9

Note: Adapted from "The performance of different propensity-score methods for estimating relative risk," by P. C. Austin, 2008, *Journal of Clinical Epidemiology*, *61*, p. 538.

In these studies, nine variables were created with varying relationships, six that were associated with the treatment assignment (x_1, x_2, x_4, x_5, x_7, x_8) and six that were associated with the outcome (x_1 through x_6). These associations were varied in magnitude to be strongly related, moderately related, or not related with either the treatment assignment or the outcome. By definition, those variables that were associated

with both the treatment mechanism and the outcome measure are true confounders (x_1, x_2, x_4, x_5). Variables x_7 and x_8 were associated only with the treatment assignment and variables x_3 and x_6 were associated only with the outcome measure. Variable x_9 was associated with neither the treatment assignment nor the outcome measure.

A series of propensity score models were derived from differing combinations of these variables for the purpose of evaluating their effectiveness in adjusting for potential bias. The primary difference in each study was the type of treatment effect that was estimated: the conditional odds ratio, hazard ratios, rate ratios (Austin et al., 2007), and relative risk (Austin, 2008). Four models that were described in these studies were of central interest as they were most like those used in observational studies. The "true propensity score model" included those six variables that were associated with the treatment assignment. A model such as this would result from research that selects variables based upon the statistical relationships between covariates and treatment assignment but not between the covariates and the outcome. The "true confounder model" consisted of those four covariates that are associated with both the treatment assignment and the outcome measure. The "potential confounders model" included all variables that were related to the outcome measure. The "full propensity model" included all nine variables (Austin, 2008). This final model is reflective of observational studies in which the variable-selection rationale might be based upon theoretical rather than statistical relationships with the treatment assignment and the outcome measure. This model might result from a study in which the researcher decided to include many variables with less consideration of theory.

The resulting estimated propensity scores were included in both greedy matching and stratification for the purpose of evaluating their performance in addressing the bias in the average treatment effect estimations. Because these studies simulated the relationships between variables, the *true* treatment effect was known and could be compared to the *estimated* treatment effect resultant from each propensity score model. They found that propensity score models that omitted variables that were related to the treatment assignment consistently resulted in biased estimations of the treatment effect (Austin et al., 2007; Austin, 2008). The precision of the effect estimation suffered when variables associated with the outcome measure were omitted (Austin, 2008).

Multilevel Data in Education

Data that are gathered from sources in educational settings that include observations beyond a single classroom are inherently multilevel in nature, because the structure of the educational system is inherently multilevel: children are nested within classrooms/teachers, classrooms are nested within schools, schools are nested within districts, etc.. Individuals who share contextual characteristics are more similar to each other than those who do not share that context. In studies that involve multiple groups, independence of observations cannot be assumed. Doing so can result in decreased standard errors for treatment effects which lead to increased Type I error rates (Raudenbush & Bryk, 2002). These particular consequences are less of a concern for estimating the propensity score, as the goal is not to accurately estimate the effects of the predictors upon the treatment assignment, but to create a successful balancing score; however, the effects of ignoring the multilevel nature of the data when estimating propensity scores can result in biased matches.

Rosenbaum (1986) recognized the challenge of multilevel data in educational research in his study of the effects of dropping out of high school on academic achievement. In this study, he matched youth who dropped out of high school with those who stayed in high school based upon their pretreatment covariates through an estimated propensity score. Rosenbaum recognized that economically disadvantaged communities had greater drop-out rates than those in wealthier communities. By ignoring community effects on drop-out rates, his analysis would tend to compare students who dropped out of disadvantaged schools to students who remained in wealthier schools.

Rosenbaum (1986) offered two potential courses of action to address the multilevel nature of the data. First, he suggested that the propensity scores could be estimated from a logistic regression model containing the relevant student covariates and a binary covariate for each school. This model would result in successful balancing on each of the individual level predictors as well as the school-membership indicators. Such a method would successfully balance any covariate that was constant within a school, whether observed or unobserved. One challenge of this model was that, with the addition of each school, the degrees of freedom would decrease. Although this is less of a concern with propensity score estimation, Hong (2004) argued that, with insufficient numbers of treatment group members per school, as might be the case when the treatment was rare, insufficient degrees of freedom would be available to estimate the fixed effect of each school upon the treatment assignment. As previously noted, the data set that was the foundation of Rosenbaum's study included 1,015 schools, a number which he deemed too large to include in the propensity score estimation. Such a method would be limited to studies that included smaller numbers of clusters.

The final model that Rosenbaum (1986) used to implement his study of the effects of dropping out of high school did not include any school membership variable in the propensity score estimations. He addressed the effect of the schools through limiting matches of treatment and control group members to within schools. By selecting this method, he argued that the between-school components of the variability were successfully controlled. Average treatment effects could then be estimated without accounting for the clustering effects of the data in the propensity score. An additional benefit of this method is that cluster-level characteristics that affected treatment assignment would be balanced within clusters regardless of whether or not they were measured. A limitation of this method, however, is that a close match for each treated individual is not always available within each cluster (Hong, 2004).

In order to overcome the requirement to limit matches to within clusters, Hong (2004) argued that relevant cluster-level confounders should be included in the propensity score estimation model. A model that included all relevant confounders at both the individual level and cluster level would allow matching across clusters because the selection bias associated with treatment assignment would be effectively controlled, rendering the treatment assignment ignorable in average treatment effect estimations. Comparing the predicted treatment assignment based upon the included pretreatment covariates to the actual treatment assignment can test this strong ignorability of treatment assignment.

Hong and Raudenbush (2005, 2006) modeled treatment assignment in a multilevel setting using a hierarchical logistic model with random school effects and fixed slopes for the individual-level and cluster-level predictors. With this model, balance

was assumed in the distribution of the individual-level pretreatment covariates, the cluster-level pretreatment covariates, and the random cluster effects. Additionally, by including the random cluster effects, they argued that any cluster-level predictors that were absent from the propensity score estimation model would be addressed through the removal of the remaining selection bias associated with the cluster. Because strong ignorability of the treatment assignment mechanism was assumed given the model, between schools comparisons were justified. Because of these conditions, the researchers chose to stratify the sample on the propensity score.

In their study, the decision to include specific variables in the propensity score estimation model was primarily based upon statistical significance. Hong and Raudenbush (2005) included those variables at both the individual- and cluster-level that showed a significant bivariate relationship with the treatment assignment. By fixing the slopes of individual-level predictors, Hong and Raudenbush (2005, 2006) assumed that the effect of those predictors upon the treatment assignment were constant across schools. The influence of contextual factors upon the treatment assignment was, therefore, limited to the intercepts. Kim and Seltzer (2007) argued that assuming that these slopes are fixed is problematic when cross-level interactions are present. An example of such a situation is when grades play a greater or lesser role in qualifying a student for an academic-improvement program based upon the funding available to the school.

Multilevel Propensity Score Estimation with Random Intercepts and Slopes. For the purposes of illustration, consider the previously stated experiment in which a researcher desires to estimate the effects of students' use of computers in the classroom on their academic achievement. The students who use computers are designated as being

in the treatment group (*Z*=1) and those who do not use computers are designated as the control group (*Z*=0). Under these conditions, the probability that the student will use computers can be modeled based upon individual- and school-level predictors. The probability that $Z = 1$ is given by p; subsequently, the probability that $Z = 0$ is given by the expression $1 - p$.

Consider Equation 1.8, which illustrates a propensity score estimation in which the intercept and slopes of the covariates are considered fixed across clusters with no cluster-level predictors and one individual-level predictor

$$logit(Z_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} \tag{1.8}$$

$$\beta_{0j} = \gamma_{00}$$

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

Or in combined form:

$$logit(Z_{ij}) = \gamma_{00} + \gamma_{10}X_{1ij}.$$

In this equation, $logit(Z_{ij})$ represents the propensity score, or the logit of the conditional probability of individual *i* in cluster *j* of receiving treatment, or, continuing with the previous example, using a computer in the classroom; β_{0j} is the average propensity score for cluster *j*; β_{1j} is the unique contribution of covariate X_1 for individual *i* in cluster *j* to the propensity score. Returning to the previous example, if the age of the student was entered into the equation as X_1, a positive estimated value for β_{1j} would indicate that, as youths grew older, the likelihood that they used computers would also increase. The parameter γ_{00} represents the mean propensity score across clusters; γ_{10} is the mean contribution across clusters of the individual-level predictor X_1 to the propensity score. Given the previous example, γ_{10} represents the mean effect of age upon

the probability of being in the treatment group across schools. The fixed intercept and slopes in this equation are modeled by not including cluster-level residuals when predicting β_{0j} and β_{1j}. Such a model is similar to a single-level model, because all clusters are modeled with the same intercept (e.g., all clusters have the same influence upon the individual's probability of receiving treatment) and the same effect upon the slope (e.g., the influence of the individual-level covariate is constant across clusters).

One variable that is not included is this multilevel model that is typically present in other linear models is the individual-level residual. The reason for its absence is due to the definition of the outcome measure when using a dichotomous outcome measure. As noted previously, the probability of being in the treatment group is given by *p*, and the probability of being in the control group is given by 1- *p*. The mean of *Z* is also equal to *p*. The variance of Z is calculated as: $p(1-p)$. The variance at the individual-level, therefore, is determined by the mean and is not a free parameter. The variance is included in the distribution of the outcome; therefore, it is not included explicitly in the individual-level model (Snijders & Bosker, 1999).

Propensity scores that are estimated from this multilevel model with fixed intercepts and slopes can be incorporated into a matching algorithm. The distance between two matched individuals can be quantified as the difference between their propensity scores. Following Equation 1.9, this difference in their propensity scores is given by:

$$PS_i - PS_{i*} = (\gamma_{00} + \gamma_{10}X_{1ij}) - (\gamma_{00} + \gamma_{10}X_{1i*j}) = \gamma_{10}(X_{1ij} - X_{1i*j}) \qquad (1.9)$$

Equation 1.10 shows that the difference in the propensity scores between treated individual *i* and control individual *i** is equal to the difference between their covariates.

In a condition in which both slopes and intercepts are fixed across clusters, matching on the propensity score results in equal distributions of covariates in the treatment and control groups.

This same logic can be applied to the situation in which individuals are matched within clusters when slopes are fixed but the intercepts are allowed to vary. By allowing intercepts to vary across clusters, the deviation of each cluster from the mean probability of receiving treatment can be modeled. This deviation is quantified through the cluster-level residual, u_{0j}. In this case, the propensity score estimation would be conducted according to Equation 1.10, below:

$$logit(Z_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} \tag{1.10}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

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

Or in combined form:

$$logit(Z_{ij}) = \gamma_{00} + \gamma_{10}X_{1ij} + u_{0j}$$

In this case the difference in propensity scores between treatment individual, *i*, and control individual, *i**, who are members of the same cluster, *j*, is given by:

$$PS_i - PS_{i*} = (\gamma_{00} + \gamma_{10}X_{1ij} + u_{0j}) - (\gamma_{00} + \gamma_{10}X_{1i*j} + u_{0j}) \tag{1.11}$$

$$= \gamma_{10}(X_{1ij} - X_{1i*j})$$

By matching within clusters, the cluster-level random effects for the treatment and control individuals, u_{0j}, cancel, and the resulting difference between two individuals on their propensity score is equal to differences in their covariates. When slopes are fixed and intercepts are allowed to vary across clusters, matching on the propensity score

within clusters effectively creates equality between treatment and control individuals on those pretreatment covariates included in the propensity score estimations.

Including cluster-level predictors of the intercept to this situation does not affect the performance of the propensity scores in balancing covariates when matching within clusters. Equation 1.12 shows a model in which a single cluster-level predictor of the intercept has been added to the model to explain the deviation of the cluster from the mean probability of receiving treatment. The intercept is allowed to vary across clusters and the slope of the individual-level predictor remains fixed across clusters

$$logit(Z_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} \tag{1.12}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

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

Or in combined form:

$$logit(Z_{ij}) = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{1ij} + u_{0j} \quad .$$

The difference in propensity scores between treatment individual, *i*, and control individual, *i**, who are members of the same cluster, *j*, is then given by:

$$PS_i - PS_{i*} = (\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{1ij} + u_{0j}) - (\gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{1i*j} + u_{0j}) \tag{1.13}$$

$$= \gamma_{10}(X_{1ij} - X_{1i*j})$$

Given the situation in which cluster-level predictors, W_j, are included, matching within clusters results in the cancelling of the effects of those variables on the propensity score difference. The difference in propensity scores between matched pairs within clusters is equivalent to the difference in the individual-level covariates.

In Equation 1.14, a cluster-level predictor has been added and the slope of the individual-level predictor is allowed to vary, as indicated by the inclusion of the cluster residual, u_{1j}.

$$logit(Z_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} \quad (1.14)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \quad .$$

Or in combined form:

$$logit(Z_{ij}) = \gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{1ij} + u_{0j}.$$

Given this equation, the difference in two individuals' propensity scores who are in the same cluster is:

$$PS_i - PS_{i*} = [\gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{1ij} + u_{0j}] \quad (1.15)$$

$$-[\gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{1i*j} + u_{0j}] \quad .$$

$$= (\gamma_{10} + \gamma_{11}W_j + u_{1j})(X_{1ij} - X_{1i*j})$$

The difference in propensity scores for the two individuals is now a function of both the individual-level characteristics and the cluster-level characteristics. The propensity scores that are estimated through this model may shift from those estimated through the model with fixed slopes, but, because the value for $(\gamma_{10} + \gamma_{11}W_j + u_{1j})$ is constant within a cluster. If individuals are sorted in ascending order based upon their propensity scores, the order of the propensity scores will not change within each cluster regardless of which of the previous models are used; therefore, when conducting within-cluster matching, the matched pairs will not change per equation.

Finally, consider the case of a propensity score estimation that includes multiple individual-level predictors that are allowed to vary across clusters and one cluster-level predictor as given in Equation 1.16:

$$logit(Z_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} \quad (1.16)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}W_j + u_{2j}.$$

Or in combined form:

$$logit(Z_{ij}) = \gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{1ij} + (\gamma_{20} + \gamma_{21}W_j + u_{2j})X_{2ij} + u_{0j}.$$

Given this equation, the difference in two individuals' propensity scores who are in the same cluster is:

$$PS_i - PS_{i*} = (\gamma_{10} + \gamma_{11}W_j + u_{1j})(X_{1ij} - X_{1i*j}) + (\gamma_{20} + \gamma_{21}W_j + u_{2j})(X_{2ij} - X_{2i*j}). \quad (1.17)$$

If removing the cluster-level predictors from this equation, the following difference is derived:

$$PS_i - PS_{i*} = \gamma_{10}(X_{1ij} - X_{1i*j}) + \gamma_{20}(X_{2ij} - X_{2i*j}). \quad (1.18)$$

Kim and Seltzer (2007) argue that these two differences can be conceptualized as "weighted combinations of differences in characteristics" (p. 9) of two individuals, *i* and *i**. When a cluster-level predictor places more weight upon the relationship between one individual-level predictor and less weight upon that relationship with a second individual-level predictor, the sorted order of the propensity scores within each cluster might

change. This change in order can result in different matches between control and treatment group members.

Kim and Seltzer (2007) tested this argument by exploring the performance of four propensity score models in balancing covariates between treatment and control individuals in a multilevel setting using within cluster matching. They based their study upon a real data set, rather than simulated data. In the first model, both slopes and intercepts were allowed to vary and both individual and cluster-level predictors were included ($Z = 1|x, w, u_i, u_s$), where x represents individual-level predictors, w represents cluster-level predictors, u_i represents random intercepts, and u_s represents random slopes. A reduced form of this model was also evaluated in the second model in which no cluster-level predictors were included in the estimation of the propensity scores ($Z = 1|x, u_i, u_s$). A third model was incorporated in order to explore the effectiveness of Hong and Raudenbush's (2005) fixed slopes propensity score estimation and included cluster-level predictors only for the intercept ($Z = 1|x, w, u_i$). The fourth model was a simple single-level propensity score estimation, including only individual-level predictors ($Z = 1|x$). Once the propensity scores were estimated, a greedy matching algorithm was used; however, matching was limited to within programs in order to maintain the multilevel structure of the data after matching.

The researchers included a number of performance indicators for each propensity score model. The mean distance between the propensity scores of the matched pairs was determined, with smaller distances indicating superior performance. The models that included random intercepts and slopes outperformed the models that did not allow slopes

to vary randomly; however, whether or not the level-2 predictors were included in modeling the slopes had no effect upon this mean distance.

Kim and Seltzer (2007) also evaluated the balance accomplished on the pretreatment covariates across the sample and within each cluster. This balance was assessed through simple descriptive and statistical mean comparisons. In addition, the balance across and within clusters was determined through the application of the multilevel model presented in Equation 1.19. In addition, this model allowed a measure of the between-cluster variation.

$$X_{ij} = \beta_{0j} + \beta_{1j}Z_{ij} + r_{ij} \tag{1.19}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

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

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right)$$

Kim and Seltzer found that all models performed well in balancing covariates across clusters as indicated by non-significant values of γ_{10}. The models that included random intercepts and slopes, however, resulted in samples with smaller values of τ_{11} indicating that balance in individual clusters was better maintained. Again, no difference was found in the balance accomplished using the two models that allowed slopes to vary randomly. The two models that included random slopes outperformed the model with fixed slopes and the single-level model to the greatest degree in clusters that showed evidence of the greatest initial imbalances.

Between-Cluster Matching using Propensity Scores

Hong and Raudenbush (2005, 2006) argued that between-cluster matching using an estimated propensity score could be used to remove bias associated with the treatment

assignment mechanism in a multilevel setting if those predictors associated with treatment assignment at the individual-level and the cluster-level were included in the estimation of the propensity score. Kim and Seltzer (2007) offered suggestions for between-cluster matching that maintained the multilevel structure of the data, stating that matches for treatment group individuals should first be made within-clusters, then between-cluster matches could be made. However, given Kim and Seltzer's argument regarding the effects of cluster-level predictors upon the propensity score estimations, the effectiveness of propensity scores as balancing scores when estimated from multilevel equations is as yet untested using simulation techniques.

The conditions apparent in the Hong and Raudenbush (2005, 2006) and Kim and Seltzer (2007) studies offer some suggestions for effectiveness of these propensity scores given the conditions of their data sets. Questions related to sample size, treatment and control ratios per cluster, and the differing relationships between cluster-level predictors and treatment assignment upon the effectiveness of multilevel propensity score estimations are but a few questions to address. Both Hong and Raudenbush (2005, 2006) and Kim and Seltzer (2007) used large data sets in their studies, with *N*s equal to 10,726 and 17,234, respectively. Other aspects of their datasets were widely varying, as illustrated in Table 2. Although the overall sample sizes and the ratio of treatment group members to control group members per cluster are similar, the sample characteristics are quite different. The effects of these sample differences on the performance of estimated propensity scores are unknown.

Table 2:

Sample Characteristics of Datasets from Multilevel Observational Studies

Sample Characteristics	Hong & Raudenbush (2005, 2006)	Kim & Seltzer (2007)
Individual-level sample size	10,726	17,234
Cluster-level sample size	1,080	29
Treatment group sample size	471	1,461
Control group sample size	10,255	15,773
Average sample size per cluster	10	594
Average number of treatment group members per cluster	0.5	50
Average number of control group members per cluster	9.5	544
Ratio of Treatment to Control group members	1:10	1:11

Problem Statement

The current study has two central purposes focused upon filling in the knowledge gaps regarding the application of propensity scores for balancing pretreatment covariates in multilevel settings. This study addresses these purposes by first exploring the performance of propensity scores estimated from data with a multilevel structure in attaining balance on covariates between treatment and comparison groups. Sample characteristics that were of specific interest are the individual-level and cluster-level

sample sizes and the ratio of treatment to control group members. Second, the performance of multilevel propensity scores using the common techniques of greedy matching within clusters, greedy matching across clusters, and quintile stratification were explored. Previous research using multilevel propensity scores have incorporated existing observational data in which the relationships among variables were uncontrolled. Subsequently, the generalizability of these studies to other research endeavors is unknown. Through examining the performance of multilevel propensity scores estimated through a Monte Carlo simulation, these conditions were controlled and the generalizeability of the findings increased. Finally, in order to examine the applicability of the findings from the simulated conditions to those apparent in collected educational data where these propensity scores might be applied, multilevel propensity scores were estimated from data available from the ELS:2002 (NCES, 2004) and their performance evaluated for the successful balancing of covariates.

CHAPTER 2

METHOD

A Monte Carlo simulation study was conducted in order to explore the extent to which propensity score methods removed selection bias in a multilevel setting as evidenced by covariate balance. Each replication involved four steps: (a) sample creation; (b) propensity score estimation from the sample; (c) utilization of the propensity scores to remove selection bias; and (d) evaluation of the performance of the propensity scores. As a follow-up study, propensity score methods were applied to a large-scale national survey that has a multilevel data structure. The applied study was conducted in order to offer insight into the performance of propensity scores in a multilevel context when that context was either ignored or included in the estimations of the propensity scores.

The Simulation

Monte Carlo simulations were used to examine the performance of propensity scores for reducing covariate bias between treatment and control groups in multilevel settings. Five independent variables were manipulated for the purpose of exploring the performance of the propensity scores: (1) level-1 and level-2 sample size; (2) ratio of treatment to control group members within clusters, (3) correlation of level-1 and level-2 predictors with treatment assignment, (4) inclusion of predictors for propensity score estimation, (5) and method of propensity score utilization to address selection bias.

Data Creation

Sample characteristics. A total of 27 samples were simulated as described in Table 3 below. The lower boundaries for the sample size were derived from conditions found in applied studies that use multilevel estimations of propensity scores. The sample

used by Hong and Raudenbush (2005, 2006) included a national dataset with an average level-1 sample size of 10 individuals per cluster and a cluster sample size of over 1,000 schools. Kim and Seltzer's (2007) applied study had an average level-1 sample size of more than 500 individuals per school and a level-2 sample size of 30. Based on these studies, the smallest sample sizes for the simulated samples were 10 individuals per cluster and 30 clusters. The larges sample sizes in these applied studies were over 500 at level-1 and over 1,000 at level-2. The very large sample sizes that would result if these conditions were simulated would be relatively uninformative due to the ease of matching with such a large pool of controls. A more informative sample size incorporated in this study was 50 individuals at level-1 nested within 100 clusters. This upper boundary for the level-1 sample size is similar to those used in simulation studies that explored sample sizes in multilevel data (Estes, 2008; Hox & Maas, 2006); the upper boundary for the level-2 sample size is applicable to those available in national datasets to which propensity score methodology is often applied. In order better to distinguish critical characteristics of the sample in the performance of propensity scores applied to multilevel data, a moderate sample size condition was also applied with a level-1 sample size of 30 and level-2 sample size of 50.

Three ratios of treatment to control group members were selected in order to simulate the ease of matching of control individuals to treatment individuals. Conditions that contribute to easy matching are those that have large numbers of control group members compared to treatment group members. A 1:1 ratio was selected based upon the definition of hard-matching by Gu and Rosenbaum (1993). A 1:9 ratio was chosen because it was similar to those in the applied studies of Hong and Raudenbush (2005,

2006) that were 1:20 and to the study by Kim and Seltzer (2007) that was 1:10. The choice of the 1:3 ratio was selected to represent moderate matching difficulty and was also used by Gu and Rosenbaum (1993). Given a total sample size of 100, these ratio conditions would equal to control units numbering 90, 75, and 50 for ratios of 1:9, 1:3, and 1:1, respectively.

Table 3:

Description of Simulated Samples

Level-1 Sample Size	Level 2 Sample Size	Treatment: Control Ratio	Total Sample Size
10	30	1:1, 1:3, 1:9	300
	50	1:1, 1:3, 1:9	500
	100	1:1, 1:3, 1:9	1,000
30	30	1:1, 1:3, 1:9	900
	50	1:1, 1:3, 1:9	1,500
	100	1:1, 1:3, 1:9	3,000
50	30	1:1, 1:3, 1:9	1500
	50	1:1, 1:3, 1:9	2,500
	100	1:1, 1:3, 1:9	5,000

In order to create the given treatment to control ratio conditions, first, a sample size that exceeded the largest sample size for each level-1 by level-2 crossed condition was created. From this sample, individuals were randomly removed from the treatment or

control groups per cluster until the desired ratio of treatment to control groups was attained.

Variable Definitions. In a series of simulations (Austin et al., 2007; Austin, 2008), the performance of several propensity scores that were estimated by including selected covariates was evaluated. In these studies, the propensity scores were estimated by including variables that were related to the treatment assignment mechanism, to the outcome, or to both. The strength of these relationships was also varied, as was the strength of the treatment effect. The current study adapted the conditions described in these studies with two essential modifications: (1) relationships of predictors to an outcome measure were not considered and (2) predictors were included at both level-1and level-2. The data generation process was conducted via the following multilevel equation:

$$logit(Z_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} \quad (2.1)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_1 + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_1 + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}W_1 + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}W_1 + u_{3j}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ \tau_{10} & \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{20} & \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix} \right).$$

In this multilevel equation, $Logit(Z_{ij})$ is the propensity score, or the logit of the conditional probability of individual *i* in cluster *j* of receiving treatment; β_{0j} is the adjusted average propensity score for cluster *j* after controlling for the other level-1

predictors; β_{kj} is the change in the propensity score per unit increase in X_k for individuals in cluster *j*; γ_{00} is the adjusted mean propensity score across clusters, after controlling for all predictors; γ_{01} is the change in the predicted propensity score per unit increase in W_1 holding all other predictors constant; u_{0j} is the residual for cluster *j* of the propensity score; γ_{k0} is the change in predicted propensity score per unit increase in X_k, holding all other predictors constant; γ_{k1} is the change in the slope of X_k per unit increase in W_1; and u_{kj} is the residual for slope *k* in cluster *j*. The variances and covariances among the residuals are given by τ_{00} through τ_{33}.

Three continuous level-1 predictors, X_1, X_2, and X_3, were generated with a strong, moderate, or null relationship, respectively, with the treatment assignment. These three predictors were correlated with the logit of the treatment assignment, *Z*, so that ρ_{xz} was equal to 0.3, 0.2, and 0, representing a strong, moderate, and null relationship, respectively. These correlations were average correlations across clusters. When a cross-level interaction was present, the correlation varied from one cluster to the next, but the mean correlation of the sample as a whole was maintained as described above.

One continuous level-2 predictor was simulated. The correlation of the level-2 predictor and the logit of the treatment assignment was held constant so that ρ_{wz} was equal to 0.3. This correlation ensured that the cluster-membership had an effect upon the treatment assignment mechanism that can be identified even in small sample sizes, but not be strong enough to obscure the effects of the other predictors that were manipulated through this study. In order to prevent the results of this study from being influenced by multicollinearity among the predictors, all predictors were simulated so as to be

uncorrelated with one other. The conditional ICCs for the slopes of each of the three level-1 predictors and the intercept for the one level-2 predictor were held constant at 0.1.

As noted previously, the following parameters were held constant across simulated samples: the correlations of the level-1 predictors with $logit(Z_{ij})$, the correlation of the level-2 predictor with $logit(Z_{ij})$, and the ICCs. The manipulated parameter was the predictor-criterion correlations. These correlations were manipulated so as to mimic potential multilevel conditions of predictors with treatment assignment in which a moderate cross-level interaction exists for all level-1 predictors, a small cross-level interaction exists for all level-1 predictors, no cross-level interactions exists for each for the level-1 predictors related to the level-2 predictor. In the sample resultant from Generating Model C, clustering was included through ρ_{wz} only. These predictor-criterion correlations are presented in Table 4. The samples were generated in SAS 9.1 (SAS Institute, 2003). Parameter estimation was also conducted in SAS 9.1, using PROC GLIMMIX and PROC MIXED. For each cell, 1000 replications were simulated.

Table 4:

Predictor Correlations with the Logit of the Treatment Assignment

Model	x_1	x_2	x_3	W	Wx_1	Wx_2	Wx_3
Moderate	0.3	0.2	0	0.3	0.3	0.3	0.3
Small	0.3	0.2	0	0.3	0.2	0.2	0.2
None	0.3	0.2	0	0.3	0	0	0

Each of the previously described design conditions was crossed. The sample characteristics consist of three level-1 sample sizes, three level-2 sample sizes, and three treatment group to control group sample size ratios. The crossing of these sample characteristics resulted in 27 sample conditions. The relationships among the predictors and the treatment assignment were manipulated so that there were three combinations, which results in a total of 81 conditions.

Propensity Score Models

In order to assess the performance of propensity scores for balancing covariates in a multilevel context, three propensity score estimation models were examined for each of the previously described samples. First, propensity scores were estimated through a hierarchical linear model as described in Equation 2.1. The multilevel model included all predictors and their cross-level interactions as described in Table 4. These propensity scores were estimated using the following SAS 9.1 (SAS Institute, 2003) program code:

```
PROC GLIMMIX DATA= data NOCLPRINT NOITPRINT;CLASS SITE;
MODEL Z (event = '1')= X1 X2 X3 W1 WX1 WX2 WX3/DDFM=RESIDUAL SOLUTION;
RANDOM intercept X1 X2 X3 /SUB=SITE TYPE=UN;
OUTPUT out = ps pred = pscore; run;
```

Second, propensity scores were estimated using a single-level logistic regression equation that included the three level-1 predictors as well as the level-2 predictor that was collapsed to the individual-level, as described in Equation 2.2, below. Finally, propensity scores were estimated through a single-level logistic regression that included only the level-1 predictors with no cluster-level components, as described in Equation 2.3. Note that the subscripts designating group membership have been removed from these equations, or:

$$logit(Z_{ij}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 W \quad (2.2)$$

$$logit(Z_{ij}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad . \quad (2.3)$$

By crossing the three propensity score models with the 81 conditions, the number of cells was increased to 243.

Propensity Score Matching

The propensity scores resulting from each of these models were incorporated into procedures to balance the treatment and control groups on covariates. A greedy matching algorithm was applied to the estimated propensity scores using nearest-neighbor matching (Gu & Rosenbaum, 1993). Each individual who received treatment was matched to a control individual with the nearest available propensity score. The steps for this matching procedure are as follows: First, the members of the treatment group were randomly ordered. An acceptable range of propensity scores to be matched with the treated individual were calculated. This range, also called a caliper, is defined as 0.1 standard deviations of the sample's average propensity score above or below the propensity score of the individual. This caliper width is based upon the procedures used in Kim and Seltzer (2007). The first treated individual is selected and, based upon that

individual's propensity score, matched with the first control group individual with an acceptably close propensity score. The matched pair is removed from the pool of potential matches and placed into the quasi-experimental sample. This process is repeated until all acceptable matches for treated individuals have been made.

This matching procedure was conducted with two variations: One matched sample was be created by matching across the entire sample, and a second matched sample was created by limiting the pool of acceptable matches among the control group to those who shared a cluster with the treatment individual.

Propensity Score Stratification

Stratification on the quintiles of the propensity score is also a common procedure for removing selection bias. Rosenbaum and Rubin (1983) showed that 90% of the bias in the variables used to estimate the propensity score can be removed through stratification on the propensity score into quintiles at minimum. The steps associated with stratification are as follows: Individuals in the treatment and control groups that fall outside the area of overlap on the propensity score are removed from the sample. The sample is then stratified into five equally-sized groups based upon their propensity scores. For the purposes of this study, the results of the stratification procedure were evaluated at this point rather than used to make further adjustments to the propensity score model in order to improve the selection bias reduction. By applying the two matching procedures and the stratification procedure, the total number of cells was increased to 729.

Evaluating the Propensity Score Performance

The performance of each of the propensity score models were evaluated based upon the following criteria: covariate balance across the sample as a whole and within each cluster. The covariate balance that were resultant from each propensity score estimation model and adjustment method per sample were compared. Balance of covariates across the sample as a whole were determined using the following measures: γ_{10} and the Standardized Mean Difference. The balance in the covariates within each cluster was measured through determination of the variance in b_{1j} through the values of τ_{11}. The calculations of these measures are described in detail below.

Balance Achievement in Predictor Covariates: γ_{10} and τ_{11}

One measure of the success of the propensity score adjustment procedure in balancing covariates is through the removal of the relationship between the treatment assignment (*Z*) and each covariate. A non-significant relationship between these variables would signify the achievement of balance between the treatment and control group on the covariate. As noted in Equations 2.4 and 2.5, this relationship can be quantified through a multilevel linear model using the propensity-score-adjusted sample in which the covariate is predicted by the treatment assignment, or:

$$\begin{aligned} X_{ij} &= \beta_{0j} + \beta_{1j} Z_{ij} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} &\sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right). \end{aligned} \tag{2.4}$$

For this equation, the treatment assignment, *Z*, of individual *i* in cluster *j* is predicting each pretreatment covariate, X_{ij}. Two parameters are of specific interest: the

average difference on the covariate between the treatment and control groups, given by γ_{10}, and the between-cluster variance, given by τ_{11}. Values close to zero on both of these measures, as indicated by tests of statistical significance using an alpha-level of .05, signify the achievement of balance on the covariate across the sample as well as within each cluster (Kim & Seltzer, 2007). In order to compare the achievement of balance across conditions, the percentage of significant estimates of γ_{10} (α =.05) out of the total number of successful replications per cell was determined.

Just as the value of γ_{10} can serve as a measure of the balance on covariates across the sample as a whole, the variance in b_{1j} across clusters, given by $\tau_{11,}$ can serve as an indicator of the variation in covariate balance across the clusters. A significant τ_{11} indicates that the balance within each cluster was not achieved, even if the propensity score adjustment method successfully balanced the covariates across the sample as a whole. The significance of the τ_{11} was determined by comparing the Z value (estimate of the τ_{11} for each covariate divided by its standard error) to a critical value calculated from a normal distribution (α = .05). The percentage of significant τ_{11} per cell is discussed. Finally, the mean values of γ_{10} and τ_{11} were determined. In order to focus on the distance of the estimate from 0, the absolute values for the mean γ_{10} and mean τ_{11} are presented. For determining the balance achieved through stratification, the following multilevel equation was used:

$$X_{ij} = \beta_{0j} + \beta_{1j}Z_{ij} + \sum_{s=2}^{5} \beta_{2j}q_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (2.5)$$

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

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

In this equation, q_{ij}, $s = 2 \ldots 5$ are dummy-indicators of the propensity score strata which are fixed across schools similar to the model used by Hong (2004); γ_{10} is now interpreted as the average difference on the covariate between the treatment and control groups after controlling for the propensity score strata.

Balance Achievement in Predictor Covariates: The Standardized Mean Difference

The Standardized Mean Difference (SMD) is a measure of the difference in the representation of a covariate between two groups, in this case, between the control group and the treatment group. The SMD is given by the following equation 2.6:

$$SMD = (\bar{X}_T - \bar{X}_C)/s_p \quad (2.6)$$

Here, $\bar{X}_T - \bar{X}_C$ is the difference between the mean value of the covariate for the control group subtracted from the mean value of that covariate for the treatment group. This measure of balance appears commonly studies that incorporate propensity scores because it provides a simple statistic for determining acceptable levels of bias in the sample. The limitation of SMD is that it does not take into account the effects of the nested structure of the data. For the matching procedures, the SMD was calculated across the sample as a whole; whereas, for the stratification procedure, this value was calculated per stratum and averaged across stratum.

Finally, the resulting sample size was considered in the evaluation of the performance of each propensity score estimation model and adjustment method. Large sample loss among the treatment group can result in conclusions that are not generalizeable to the original sample. For this reason, the percentage of the treatment group that was retained after application of the adjustment method was calculated. In

addition to the percent of the treatment group retained, the percentage of the clusters retained was also determined, as loss of clusters can also contribute to reduction in external validity and decrease efficiency of parameter estimations, especially in a multilevel setting.

Application using the Educational Longitudinal Study of 2002

Description of Data. Data used in the applied portion of this study were drawn from the Education Longitudinal Study of 2002 (ELS:2002; NCES, 2004). The ELS:2002 is a nationally representative study of young people in the United States designed to monitor them from 10th grade on to post-secondary education and/or employment. In the ELS:2002, a number of instruments were used to collect data during the spring semester of the 2001-2002 academic year. Data were collected through questionnaires, test data, and observations by administration proctors. A total of 15,362 students completed the questionnaire during their 10th grade year of high school. A total of 752 schools was included in the study. The average number of students participating in the study per school was 26. In addition to student questionnaires, 13,488 parents, 7,135 teachers, 743 principals, and 718 librarians completed questionnaires. Additionally, a facilities checklist was completed by questionnaire administrators based on observations at each school. The ELS:2002 also employed a complex sampling structure in order to include a representative sample of the national population. The dataset includes weights associated with both school selection and individual selection for the purpose of generalizing to the larger population. These weights were not considered in the current research in order to maintain greater similarity to the conditions of the simulation study, although this decision did limit the generalizability of the findings from the current study.

Defining the Propensity Score Model. The ELS:2002 dataset offers a multilevel data structure that is useful for evaluating the performance of propensity scores in conditions similar to those explored through the simulation study. For the purposes of providing a context in which to evaluate the performance of multilevel propensity scores, the following research question was posed: What are the average effects of having computers in the home upon students' academic achievement in mathematics?

For this question, the treatment assignment was based upon students' responses to the following questionnaire item: "Does your family have the following in your home: A computer." Those 10th-graders who indicated "No" were designated as treatment group members and those who indicate "Yes" were designated as control group members. Although typically researchers consider the addition of something as a treatment, in the current situation, it should be noted that the treatment condition was actually a lack of something. This decision was made due to the fact that not having a computer in the home was the rarer condition. When conducting matching procedures, it is better to have a larger number of control group members from which to match treatment group members.

Description of Sample

In the initial sample, the distributions of variables between treatment groups were explored. The variables included in this initial exploration were based upon previous research (Du, Havard, Yu, & Adams, 2004) that explored the relationship of computer use and achievement in mathematics. This study incorporated linear modeling of both individual and school characteristics using data from ELS:2002. The student characteristics included as predictors of academic performance in this study were: SES,

computer use and access, race/ethnicity, educational expectations of teachers and parents, and specific mathematics courses that were taken by the 10th-graders. The school characteristics in their models included: school urbanicity and proportion of 10th-graders participating in free/reduced lunch programs.

The predictors to be included in the current study expanded those used by Du et al. (2004). Those variables that showed (1) significant imbalance between the treatment and control group as evidenced by t-test results ($\alpha = .05$) and (2) showed a significant relationship individually with the outcome measure as evidenced through OLS regression were included in the propensity score estimation model. The initial pool of variables are presented in Table 5 below. Two variables include in this list were compiled from a series of variables in order to capture constructs similar to those included in the Du et al. (2004). These two variables were "Number of math courses" and "Breadth of math courses." These two variables were compiled from 10 variables in the ELS:2002 dataset (F1S17A through F1S17J) that described whether the 10th-grader had taken a particular mathematics course. These courses were: general math, pre-algebra, algebra I, algebra II, trigonometry, pre-calculus, calculus, geometry, business math, and other math. Each variable provided a measure of the number of mathematics courses taken. In order to include the number of mathematics courses in the propensity score estimation model as a single variable, each 10th-grader's responses for these variables were summed. For the variable indicating the breadth of math courses, the number of different math courses indicated by each 10th-grader was summed.

Table 5:

Initial Differences between Treatment Group and Control Group on Selected Variables

Variable (ELS:2002 identifier)	Treatment	Control	Difference	*t-value*	*df*	*prob.*
Math number correct (BYTXMIRR)	34.566	42.401	-7.835	-10.38	1375	<.0001
Socioeconomic Status (BYSES1QU)	1.8638	2.7929	-0.929	-12.69	1375	<.0001
How often uses computer for school work (BYS45B)	2.9066	3.4393	-0.533	-7.18	1375	<.0001
How often uses computer to learn on own (BYS45C)	2.5525	3.1277	-0.575	-6.5	1375	<.0001
Student academic expectation (BYSTEXP)	4.8288	5.4795	-0.651	-7.38	1375	<.0001
Parental academic expectation (BYP81)	4.7549	5.158	-0.403	-4.6	1375	<.0001
Teacher academic expectation (BYTM20)	3.6576	4.5616	-0.904	-9.88	1375	<.0001
School Control (BYSCTRL)						
Public	0.8833	0.7857	0.0976	3.57	1375	0.0004
Catholic	0.0467	0.0848	-0.038	-2.06	1375	0.0395
Other private	0.07	0.1295	-0.059	-2.66	1375	0.0078
School urbanicity (BYURBAN)						
Suburban	0.4047	0.4268	-0.022	-0.65	1375	0.5178
Rural	0.2802	0.2795	0.0007	0.02	1375	0.9822
Urban	0.3152	0.2938	0.0214	0.68	1375	0.4985

School % 10th grade free/reduced lunch (BY10FLP)	3.7665	3.0223	0.7442	6.02	1375	<.0001
Family variables (BY10FLP)						
Biological Mother and Father	0.4553	0.6991	-0.244	-7.56	1375	<.0001
Two parents, not both biological	0.1907	0.133	0.0576	2.38	1375	0.0176
One parent	0.3424	0.1607	0.1817	6.74	1375	<.0001
Other family structure	0.0117	0.0071	0.0045	0.74	1375	0.4622
Number of math courses	7.0934	7.9482	-0.855	-5.51	1375	<.0001
Breadth of math courses	3.7665	4.2107	-0.444	-5.71	1375	<.0001
Ethnicity (F1RACE)						
Indian	0.0117	0.0018	0.0099	2.38	1375	0.0175
Asian	0.035	0.0491	-0.014	-0.97	1375	0.3336
Black	0.179	0.0598	0.1192	6.36	1375	<.0001
Latino	0.2607	0.0929	0.1678	7.5	1375	<.0001
Multi	0.0661	0.0402	0.026	1.81	1375	0.0703
White	0.4475	0.7563	-0.309	-10.07	1375	<.0001
Gender (F1SEX)						
Male	0.4202	0.4563	-0.036	-1.05	1375	0.2955
Female	0.5798	0.5438	0.036	1.05	1375	0.2955

Participant data were included in the final sample if the youth attended the same school for both the base-line year (10th grade) and the follow-up year (12th grade). Listwise deletion was used to address missing data on those variables chosen for inclusion in the propensity score model. Responses that were considered missing were

designated in the dataset as “missing,” “nonrespondent,” “multiple responses,” “don’t know,” “partial interview-breakoff,” or “survey component legitimate skip/NA.” Lastly, all observations from a school were dropped if that school did not have at least one treatment group member and one control group member.

The sample, after data cleaning, consisted of 1,377 youth and 183 schools. The treatment group consisted of 257 individuals, and the control group consisted of 1,120 individuals. This translates to a treatment-control group ratio of approximately 1:3. The average number of treatment individuals per school was 1.4 (*SD*=.77, Range: 1 - 5). The average number of control individuals per school was 6.1 (*SD*=3.9, Range: 1-17). The median treatment-control group ratio within each school was 0.25 (1:3), as was found for the sample as a whole; however, the modal ratio was 1.0 (1:1) which was present in 13% of the schools. The number of treatment and control group members per school in the final sample is described in Table 6 and Table 7 below.

Table 6:

Percent of Schools per Number of Treatment Individuals

# of Treatment Individuals	% of Schools
1	73%
2	18%
3	6%
4	3%
5	0.6%

Table 7:

Percent of Schools per Number of Control Individuals

# of Control Individuals	% of Schools
1	12%
2 to 5	35%
6 to 10	39%
11 to 17	14%

Propensity scores were estimated from each of the three models described in Equations 2.1, 2.2, and 2.3. As in the simulation study, the estimated propensity scores were incorporated into a within-school matching procedure, a between-school matching procedure, and quintile stratification. Crossing these conditions resulted in a total of nine matched samples: within-school matching using each of the three propensity score

estimation models, between-school matching using each model, and quintile stratification using each model.

Performance of Propensity Scores. As described in the simulation portion of this study, the performance of the propensity scores in each adjustment method were evaluated by determining the balance attained on the propensity scores and on each individual-level covariate across the sample as a whole and within each school using Equations 2.4, 2.5, and 2.6. The percent treatment group retained and clusters retained resultant from each procedure were also explored.

Treatment Effect Estimation. The applied portion of this investigation into the performance of propensity scores in multilevel data allowed for an exploration of the treatment effect estimations. Because the data were multilevel in nature, this effect was estimated using a multilevel model, with treatment assignment as a student-level predictor and the mathematics achievement score as the outcome measure. Both the intercept and the predictor were allowed to vary across schools. This multilevel model was applied to each of the six matched samples as well as the initial, unadjusted sample, as described by Equation 2.9.

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} Z_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} &\sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right) \end{aligned} \tag{2.9}$$

Here, Y_{ij} represents the mathematics score for student *i* in school *j*; γ_{00} is the adjusted average mathematics score across schools; γ_{10} is the average treatment effect

across schools; and e_{ij}, u_{0j}, and u_{1j} are the individual-level residual and cluster-level residuals, respectively. The resulting average effect estimations and the between-school variations in these effects (τ_{11}) for each propensity score model will be explored.

A similar model will be applied to the three stratified samples, as described in Equation 2.10, or:

$$X_{ij} = \beta_{0j} + \beta_{1j}Z_{ij} + \sum_{s=2}^{5} \beta_{2j}q_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (2.10)$$

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

$$\beta_{(2-5)j} = \gamma_{(2-5)0}.$$

In this equation, q_{ij}, $s = 2 \ldots 5$ are dummy-indicators of the propensity score strata which are fixed across schools; γ_{10} is now interpreted as the average difference on the achievement score between the treatment and control groups after controlling for the propensity score strata. The effects of the strata are fixed across schools (cf. Hong, 2004).

In order to determine the bias in the estimation of the treatment effects, a comparison between the models was conducted. First, the adjustment method by propensity score estimation model that showed evidence of the best balance across the sample and the least variation in that balance across the schools while retaining the largest percentage of the treatment group was determined. This method by model was considered to have best controlled for the selection bias compared to the other methods by models and, therefore, served as a comparison by which to determine remaining bias. The percentage of bias remaining in the effect estimates was calculated using the following equation:

$$Percent\ Bias = 100 * \left(\frac{Estimate_m - Estimate_c}{Estimate_c}\right). \quad (2.11)$$

Here, $Estimate_m$ represents the treatment effect estimate resulting from each of the model/method other than the comparison model/method; $Estimate_c$ represents the treatment effect estimate resulting from the comparison model.

CHAPTER 3

RESULTS AND DISCUSSION

Performance of Propensity Scores Estimated using a Multilevel Model

Within-cluster Matching

Balance Achievement in Predictor Covariates: Percent Significant γ_{10}

The performance of within-cluster matching using propensity scores that are estimated using a multilevel model (Model 1) is discussed in this section. Within-cluster matching is the adjustment method used by Kim and Seltzer (2005) and Rosenbaum (1986) to address imbalance in covariates for nested data. Across the 81 sample conditions, the average percent significant γ_{10} resultant from application of this method with propensity scores estimated from Model 1 for X_1 was 12.8, for X_2 was 8.1, and for X_3 was 7.2. This pattern of decreasing percent significant γ_{10} from X_1 to X_3 is parallel to the decreasing correlations in the generating population between each of these variables and the treatment assignment: imbalance was simulated to be greatest in X_1, moderate in X_2, and 0 in X_3.

Sample Size at Level 1 and Level 2. The relationship between level-1 sample size and percent significant γ_{10} was negative: as the level-1 sample size increased, the percentage of significant γ_{10} decreased. This pattern is reversed for level-2 sample size: as level-2 sample size increased, the percentage of significant γ_{10} increased (see Table 8). These relationships were similar across covariates, X_1 through X_3, although the relationship decreased as the correlation of the covariate with treatment assignment decreased.

Table 8:

Average Percentage Significant γ_{10} Resultant from Within-Cluster Matching with Propensity Scores Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	10.6	8.3	7.5	9.6	6.4	6.0	8.6	6.1	5.9	7.6
50	14.3	10.4	8.2	9.9	6.7	6.0	8.4	6.3	5.7	8.4
100	28.5	16.1	11.3	13.2	7.7	7.3	10.4	6.8	6.9	12.0
Mean	17.8	11.6	9.0	10.9	6.9	6.4	9.2	6.4	6.1	9.4

Cross-Level Interaction. The product-criterion correlations ($\rho_{(WX1)Z}$, $\rho_{(WX2)Z}$, $\rho_{(WX3)Z}$) were varied to be 0, .2, or .3 across all covariates in a given cell. The results of this study indicate that cross-level interactions do not have a bearing upon the percentage of significant γ_{10} when within-cluster matching is used. The change in the percentage of significant γ_{10} per sample condition was minimal across the cross-level interactions within each sample-size condition. The average percentages are presented in Table 9 for covariate X_1.

Table 9:

Average Percentage of Significant γ_{10} for X_1 per Cross-Level Interaction Resultant from Within-Cluster Matching with Propensity Scores Estimated using Model 1

$\rho_{(WX)Z}$	Level 1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	11.6	16.0	33.5	8.8	11.5	19.3	7.4	9.4	12.4	14.4
0.2	9.9	14.1	27.6	8.8	9.7	14.2	7.8	6.9	10.7	12.2
0.3	10.2	12.9	24.4	7.1	10.0	14.7	7.3	8.4	10.8	11.8

Treatment-Control Group Ratio. The ratio of the number of treatment group members to control group members was found to have a positive relationship with the percentage of significant γ_{10} under specific sample conditions. Findings indicate that larger differences in the number of treatment and control group members are positively related to greater percentages of significant γ_{10}. As level-2 sample size increases, this relationship becomes stronger (see Figure 1). The relationship between treatment-control group ratio and percent significant γ_{10} is consistent across level-1 sample size (see Figure 2), indicating that level-1 sample size does not have a bearing upon that relationship. The percent significant γ_{10} across all conditions resulting from within-cluster matching is presented in Figure 3. Independent variables that are represented in the axis labels in Figure3 and all similar figures are defined as follow: Labels of 1, 3, and 9, refer to the control proportion of the 1:1, 1:3, and 1:9 treatment-control ratios. Labels of 0, .2, and .3

refer to the values for the product-criterion correlations. Sample-size conditions at level-1 and level-2 are labeled.

Figure 1: Mean Percent Significant γ_{10} per Treatment-Control Ratio and Level-2 Sample Size

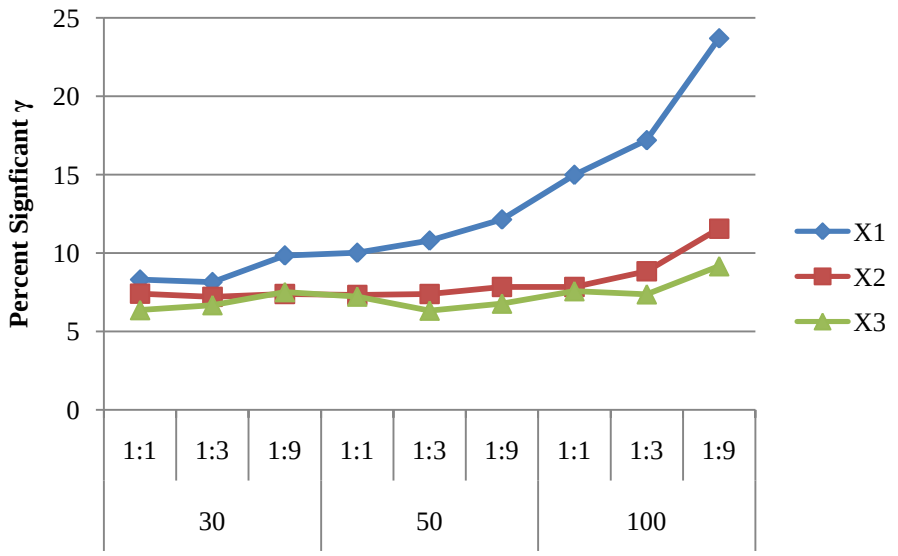


Figure 2: Mean Percent Significant γ_{10} per Treatment-Control Ratio and Level-1 Sample Size.

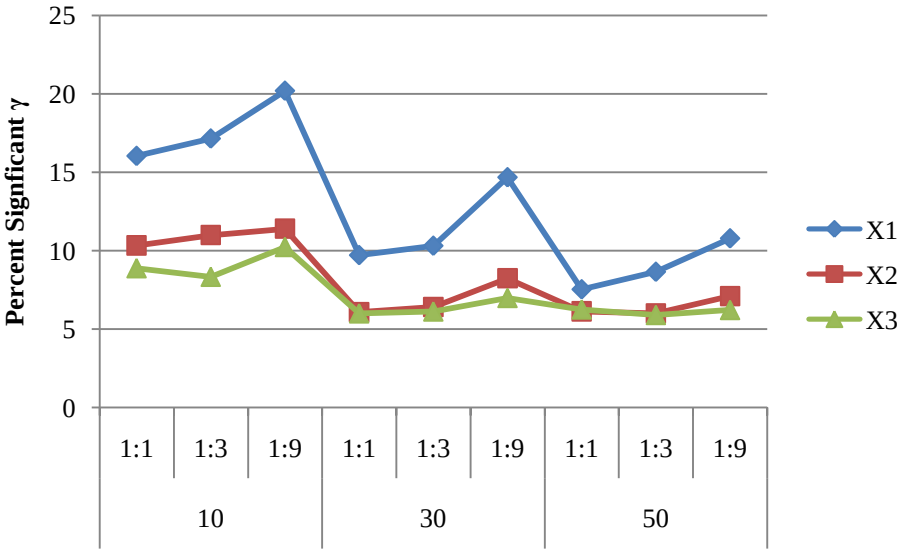
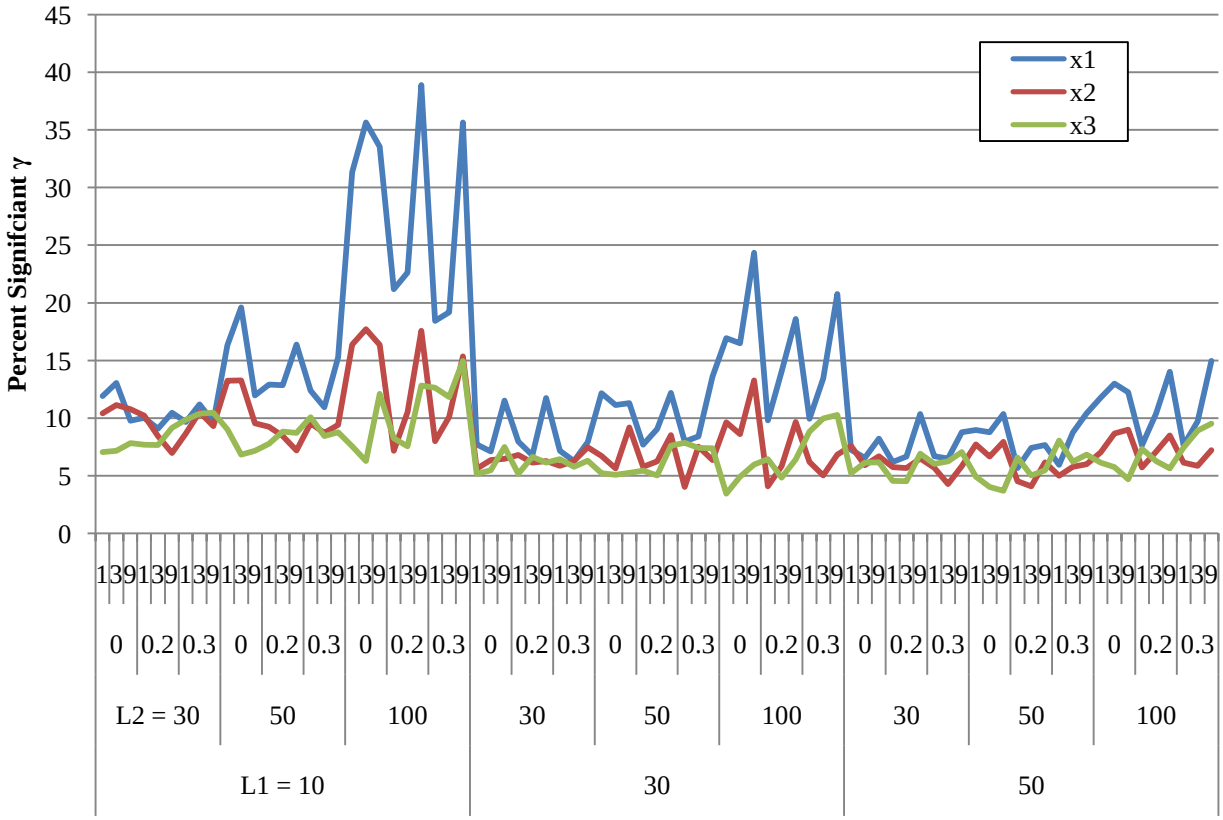


Figure 3: Percentage of Significant γ_{10} for X_1, X_2, and X_3 across Simulation Conditions for Within-Cluster Matching with a Propensity Score Estimated using Model 1



Variation in Balance Achievement within Clusters: Percent Significant τ_{11}

Sample Size at Level 1 and Level 2. A positive relationship is indicated between level-2 sample size and the percentage of significant τ_{11}. The relationship of percentage of significant τ_{11} with level-1 sample size, however, is non-linear. When sample size at level-1 is smallest (n=10) and largest (n=50), the percentage of significant τ_{11} is less than when level-1 sample size is moderate (n=30). The smallest percentage of significant τ_{11} is attained when both level-1 and level-2 sample sizes are smallest (see Table 10). The pattern is nearly identical for each covariate, X_1 through X_3.

Table 10:

Average Percentage of Significant τ_{11} Resultant from Within-Cluster Matching with Propensity Scores Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	2.1	11.5	8.8	1.8	11.4	8.6	2.0	9.9	9.0	7.2
50	6.0	14.0	11.0	5.4	13.4	10.4	5.1	13.7	10.7	10.0
100	14.3	14.3	12.9	13.4	14.9	11.8	12.1	14.3	12.4	13.4
Mean	7.5	13.3	10.9	6.9	13.2	10.3	6.4	12.7	10.7	10.2

Cross-Level Interaction

The relationship of the cross-level interaction and the percentage of significant τ_{11} is essentially 0 across each sample size condition: as the cross-level interaction increases, the percentage of significant τ_{11} remains relatively constant within each level-1 by level-2 sample size condition (see Table 11).

Table 11:

Average Percentage of Significant τ_{11} for X_1 per Cross-Level Interaction Resultant from Within-Cluster Matching with Propensity Scores Estimated using Model 1

$\rho_{(wx)z}$	Level 1 Sample Size									Mean
	10			30			50			
	Level 2 Sample Size			Level 2 Sample Size			Level 2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	3.0	5.9	13.8	11.9	15.2	15.9	9.4	11.6	14.3	11.2
0.2	1.4	5.8	14.6	10.4	13.6	14.2	7.8	10.8	12.7	10.2
0.3	1.8	6.3	14.7	12.1	13.3	12.8	9.2	10.5	11.7	10.3

Treatment-Control Group Ratio

The relationship of the percent of significant τ_{11} and treatment-control group ratio is related to the level-1 sample size. When the level-1 sample size is small (n=10), the relationship between treatment-control group ratio and percent of significant τ_{11} is negative: as the difference between the number of treatment group members and control group members increases, the percent of significant τ_{11} decreases. Once the level-1 sample size increases to 30 or 50, however, the relationship reverses: as the ratio increases, the percent of significant τ_{11} increases (see Figure 4).

The relationship of the percent of significant τ_{11} with the treatment-control group ratio is small and positive across level-2 sample-sizes, indicating no interaction effects (see Figure 5). The relationship of the percent of significant τ_{11} with the treatment-control ratio is consistently positive across cross-level interaction conditions, indicating no interaction effects (see Figure 6).

Figure 4: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-1 sample size.

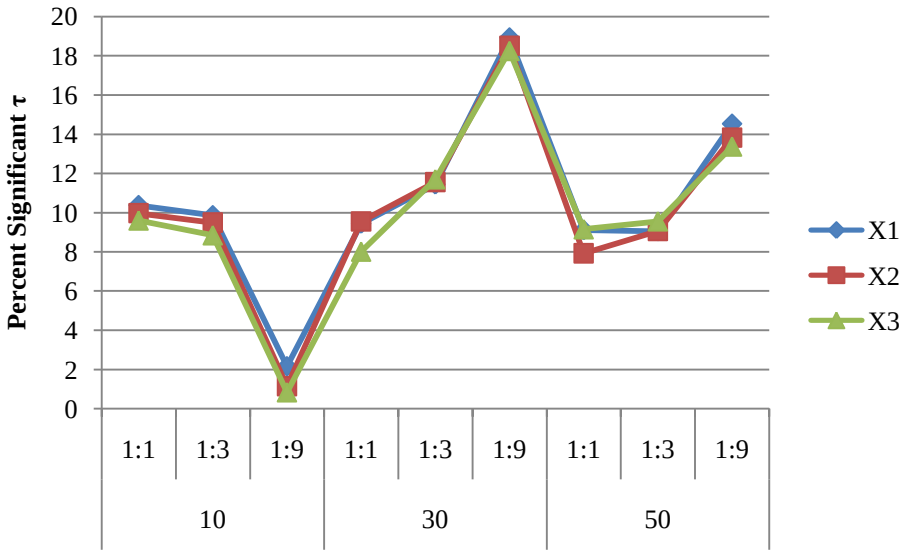


Figure 5: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-2 sample size.

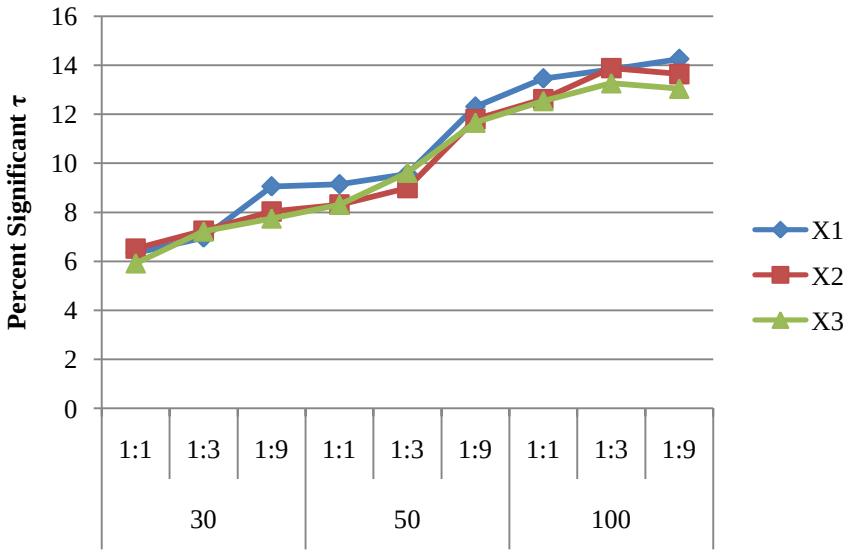
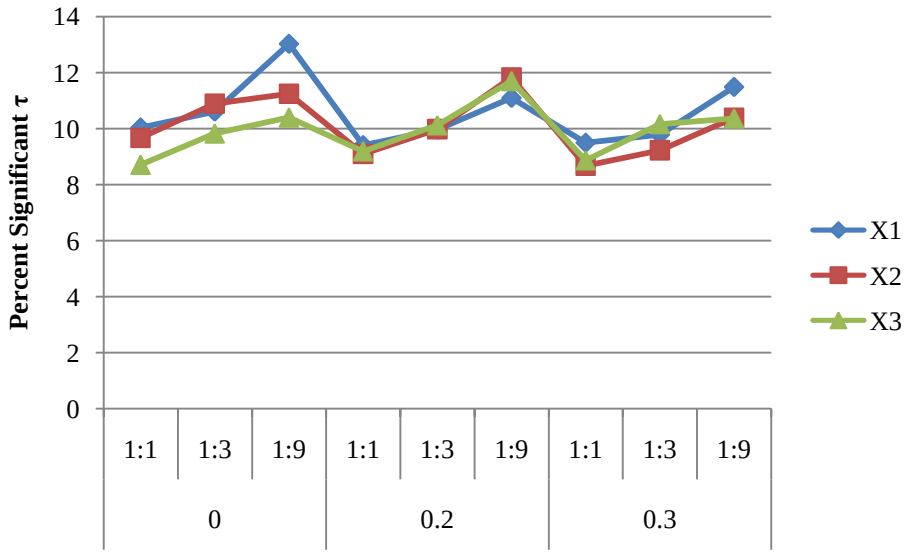


Figure 6: Mean Percentage of Significant τ_{11} per cross-level interaction and treatment-control group ratio.

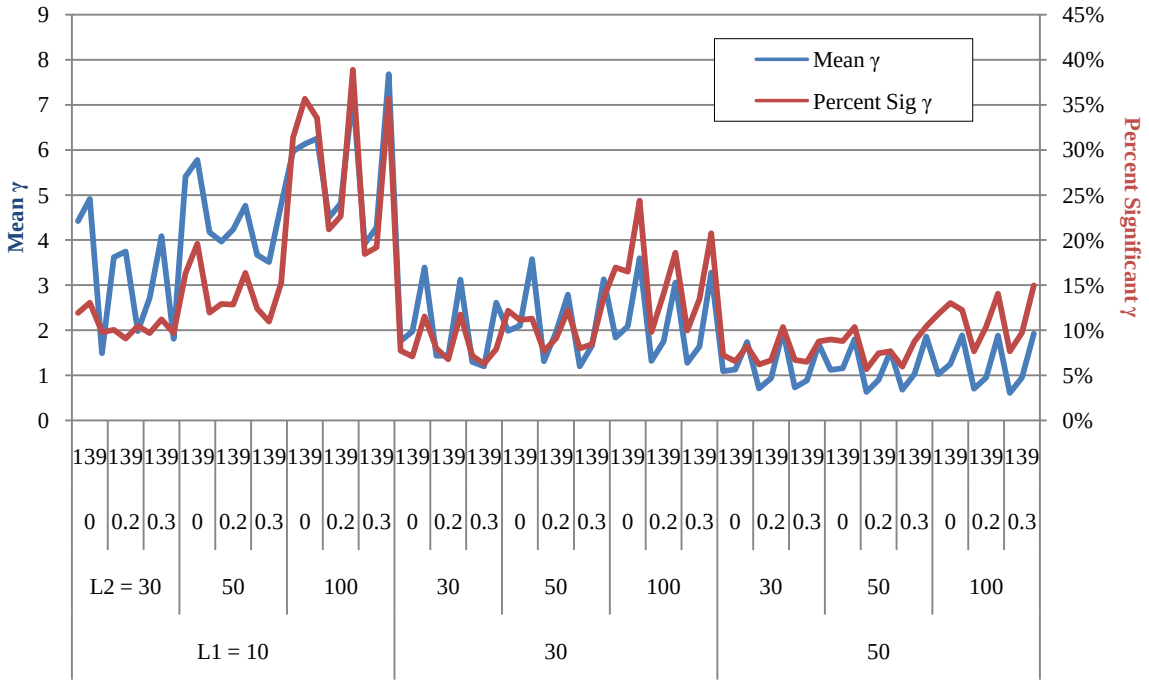


Percent Significant versus Mean Value of γ_{10} and τ_{11}

When using a propensity score that is estimated using a multilevel model and applied to within-cluster matching, the mean γ_{10} and percent significant γ_{10} are similar per condition. The balance achieved with a multilevel propensity score in within-cluster matching is least when the treatment-control group ratio is 1:9 versus 1:3 and 1:1. A plausible explanation for this finding is that the small number of treatment group members per cluster does not allow for adequate estimation of the treatment assignment mechanism that is captured by the propensity scores.

These results also indicate that balance across the sample was better accomplished when there were fewer clusters. The pattern for both percent significant γ_{10} and mean γ_{10} are similar across simulation conditions. It is apparent that the error in the propensity score estimations that are hypothesized to exist resultant from insufficient sample size at level-1 are not alleviated by increasing the number of clusters.

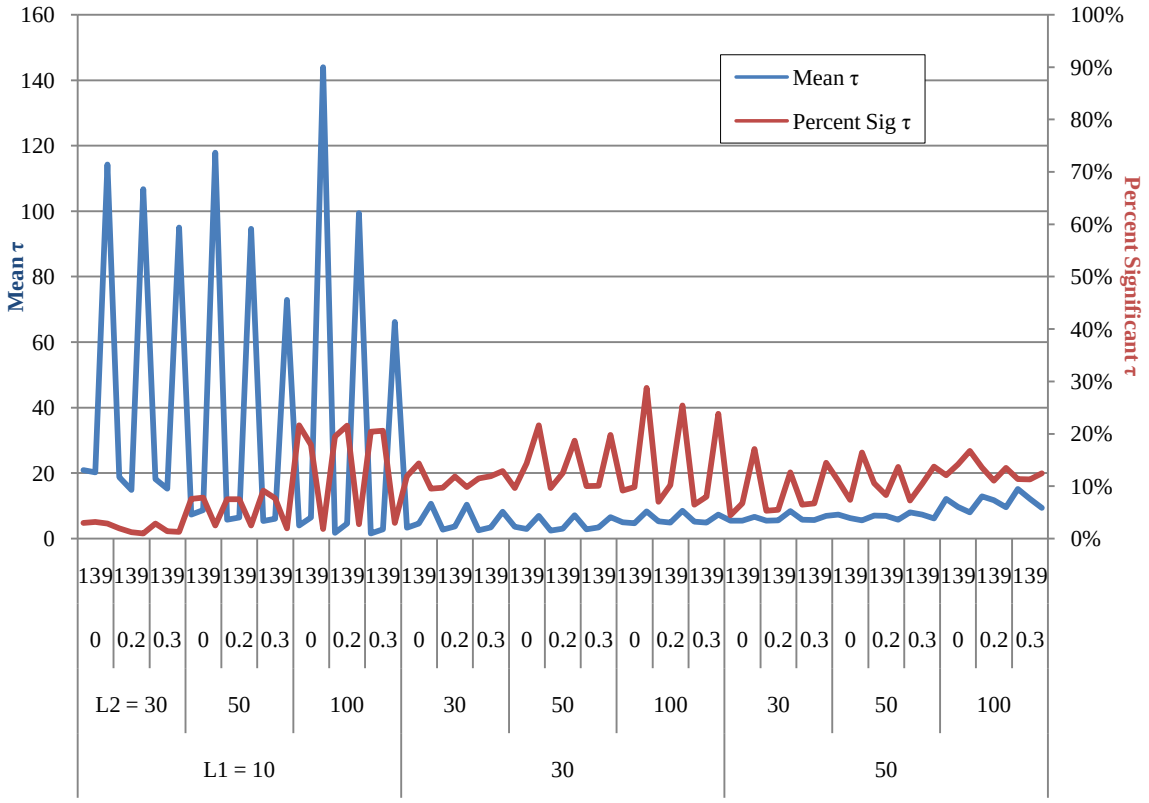
Figure 7: Overlay of Mean γ_{10} and Percent Significant γ_{10} for Within-Cluster Matching with Propensity Score Estimated using Model 1



Variance across clusters in the achieved balance is largest when sample-size at level-1 is smallest, as indicated by the fluctuations in the values of the mean γ_{10} per condition. Greater variance and smaller sample sizes decrease the power to detect differences in the slopes, which is reflected in the smaller percent significant τ_{11} when sample-size at level-1 and level-2 are smallest. Once sample size at level-1 reached 30, however, values for both the percent significant τ_{11} and the mean value of τ_{11} remain small across conditions.

Considering the findings regarding balance as indicated by the mean values of γ_{10} and τ_{11}, issues related to sample size are essential to consider when using within-cluster matching with a propensity score estimated using a multilevel model. A sufficient number of treatment group individuals must be present at level-1 in order to estimate the propensity scores adequately.

Figure 8: Overlay of Mean τ_{11} and Percent Significant τ_{11} for Within-Cluster Matching with Propensity Score Estimated using Model 1



Balance Achievement in Predictor Covariates: The Standardized Mean Difference (SMD)

Sample Size at Level 1 and Level 2. The effect of sample size upon the standardized mean difference is clear: Larger sample sizes result in smaller covariate differences between the groups. This pattern is apparent for increases in both level-1 sample size and level-2 sample size. In addition, increasing sample sizes result in less variation in the SMD across covariates, X_1, X_2, and X_3. The balance that is attained for each covariate, X_1, X_2, and X_3, is similar across conditions, indicating that within-cluster matching has almost identical results for covariates regardless of the initial strength of the

correlation of the covariates with the treatment assignment. The mean value for the SMD for each sample-size condition is presented in Table 12, below.

Table 12:

Mean SMD Resultant from Within-Cluster Matching using a Propensity Scores Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	65.3	29.5	20.5	74.7	33.8	24.1	83.1	35.5	25.8	43.6
50	53.1	24.1	16.4	57.3	26.5	18.7	59.1	26.8	19.6	33.5
100	47.0	19.9	12.9	41.9	19.1	13.7	41.2	19.6	14.3	25.5
Mean	55.2	24.5	16.6	58.0	26.5	18.8	61.1	27.3	19.9	34.2

Cross-Level Interaction. The results of this study indicate that the strength of the cross-level interactions does not have a bearing upon the SMD when applying within-cluster matching using a multilevel propensity score estimation model. The average percentages are presented in Table 13 for covariate X_1.

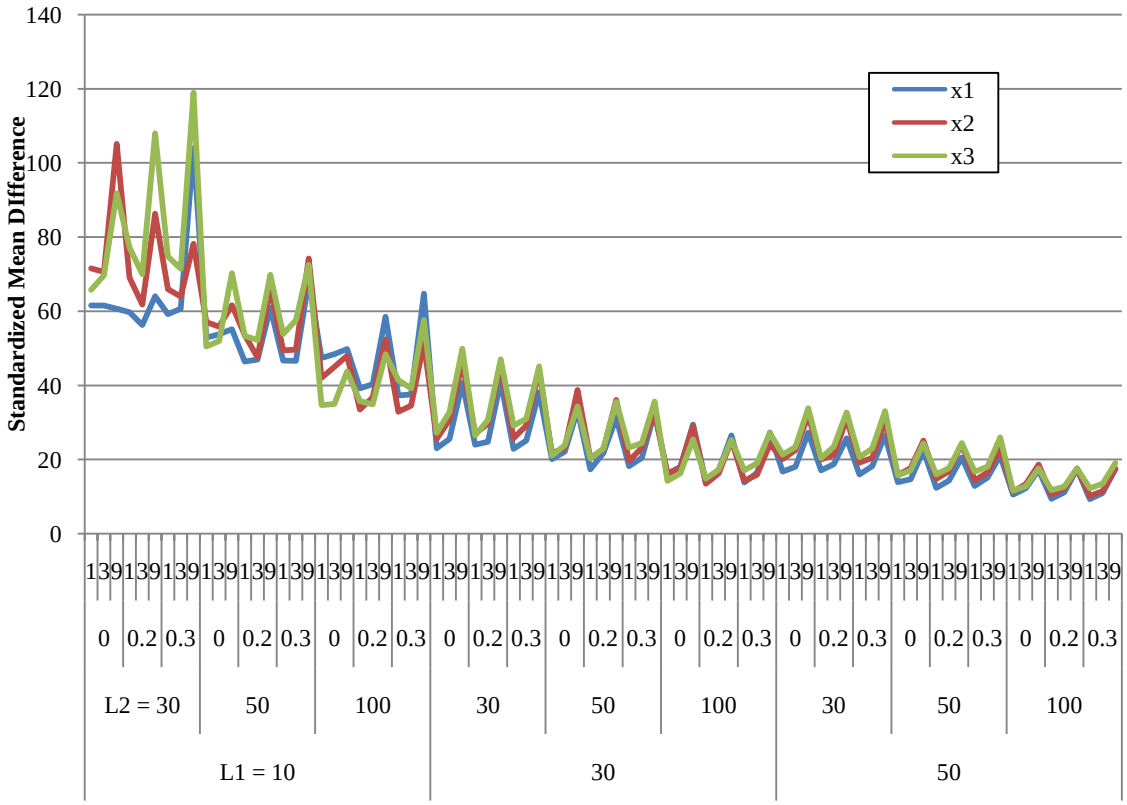
Table 13:

SMD for X_1 per Cross-Level Interaction Resultant from Within-Cluster Matching using a Propensity Score Estimated using Model 1

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	61.3	53.9	48.6	29.8	25.2	21.2	20.7	17.0	13.3	32.3
0.2	60.1	51.5	46.0	30.0	23.5	19.2	20.5	15.7	12.7	31.0
0.3	74.7	53.9	46.6	28.7	23.8	19.2	20.2	16.3	12.6	32.9

Treatment-Control Group Ratio. The relationship between the treatment-control group ratios and the SMD showed a consistent pattern across sample size conditions: The 1:9 ratio condition consistently showed larger SMDs compared to the 1:3 and 1:1 condition. Larger differences between the sizes of the treatment group versus control group members. Additionally, findings indicate that, given sample characteristics in which level-1 sample size is small and level-2 sample size is large, the relationship between the treatment-control group ratio and mean SMD is greater. The SMD for each condition is presented in Figure 9 below.

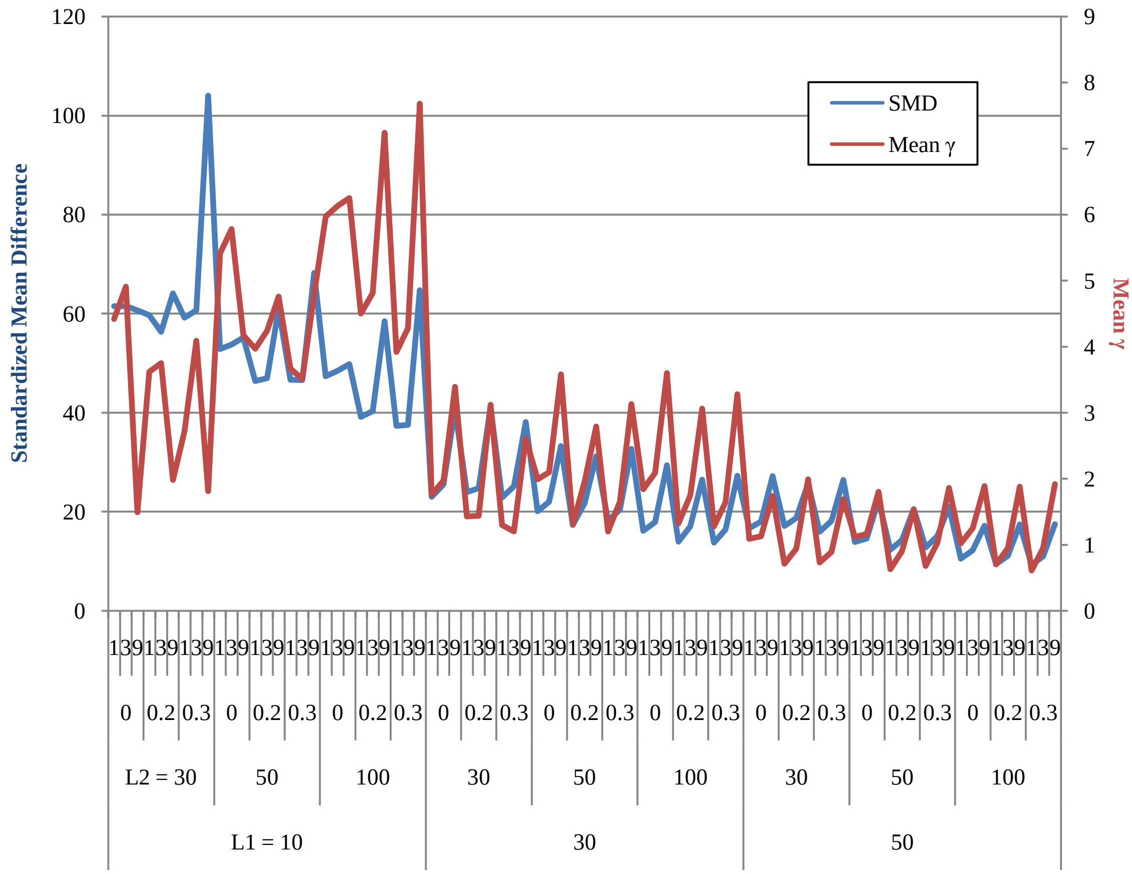
Figure 9: Standardized Mean Differences for X_1, X_2, and X_3 per Simulation Condition for Within-Cluster Matching with a Propensity Score Estimated using Model 1



SMD versus Mean γ_{10}

Across the level-1 sample size conditions, the SMD consistently decreased, while the mean γ_{10} remained relatively unchanged once level-1 sample size reached 30. Differences in these indicators of balance across the sample as a whole are apparent when level-1 sample size is equal to 10. The differences in these indicators are likely resultant from the error remaining in the estimations of γ_{10} due to the small sample size. The calculations of SMD, however, are less influenced by sample size and subsequently show a more consistent pattern across sample-size conditions.

Figure 10: Overlay of SMD and Mean γ_{10} for Within-Cluster Matching with Propensity Score Estimated using Model 1



Between-cluster matching

Balance Achievement in Predictor Covariates: Percent Significant γ_{01}

Between-cluster matching is the typical method employed in propensity score matching studies. The application of this method is usually not purposefully chosen but is the default when the clustered nature of the data is not recognized or is considered irrelevant. In this propensity score adjustment condition, nearest neighbor matching is conducted across clusters, or rather, without consideration of the clusters.

Sample Size at Level 1 and Level 2. The findings from this study indicate that the percentage of significant γ_{10} decreases as the sample-size at level-1 increases. This pattern

is evident in the final row of Table 14 for the mean γ_{10} for X_1 and somewhat for X_2; this pattern, however, is not apparent for X_3, which had the lowest percentages of significance across level-1 sample sizes. The relationship of level-2 sample size with the percent of significant γ_{10} is consistent across X and level-1 sample sizes: As the level-2 sample size increases, the percent of significant γ_{10} increases.

Table 14:

Average Percentage of Significant γ_{10} for X_1 Resultant from Between-Cluster Matching with a Propensity Score Estimated using Model 1

Level- 2 Sample Size	X_1 Level-1 Sample Size			X_2 Level-1 Sample Size			X_3 Level-1 Sample Size			Mean
	10	30	50	10	30	50	10	30	50	
30	15.0	15.3	13.6	7.4	8.4	8.4	4.8	6.3	7.1	9.6
50	28.0	24.9	19.3	12.0	10.7	10.6	5.3	7.9	8.1	14.1
100	53.8	44.1	34.4	23.6	18.1	15.5	8.5	10.2	11.2	24.4
Mean	32.2	28.1	22.4	14.3	12.4	11.5	6.2	8.1	8.8	16.0

Cross-Level Interaction. The results of this study indicate that cross-level interactions are negatively related to the percentage of significant γ_{10} when propensity scores estimated using a multilevel model are applied using between-cluster matching: as the cross-level interaction increases, the percent of significant γ_{10} decreases. Additionally, this relationship was most pronounced among the γ_{10} of X_1, the covariate which had the largest correlations with the treatment assignment, followed by X_2, and little relationship with X_3. The percentage of significant γ_{10} per sample characteristic for X_1 is presented in

Table 15 below. The final column of this table illustrates the decrease in the percentage of significant γ_{10} from the cells as the cross-level interaction increases.

Table 15:

Average Percentage of Significant γ_{10} for X_1 per Cross-Level Interaction Resultant from Between-Cluster Matching with a Propensity Score Estimated using Model 1

$\rho_{(WX)Z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	22.9	42.9	73.9	21.4	35.8	61.7	19.0	28.0	49.8	39.5
0.2	12.9	23.2	48.1	14.1	21.3	39.5	12.0	16.4	30.5	24.2
0.3	9.1	17.8	39.4	10.5	17.7	31.2	9.7	13.5	22.9	19.1

Treatment-Control Group Ratio

The ratio of treatment group members to control group members has a bearing upon the percentage of significant γ_{10} under specific conditions. The finding for between-cluster matching is opposite that indicated in the previously discussed condition, within-cluster matching. In the current condition, findings indicate that larger differences between the number of treatment and control group members within each cluster are negatively related to the percentages of significant γ_{10}: As the number of control individuals becomes greater relative to the number of treatment individuals, the percentage of significant γ_{10} decreases. Additionally, the relationship of the treatment-control group ratio and the percentage of significant γ_{10} is more pronounced when level-2

sample increases (see Figure 11) whereas it remains nearly constant across level-1 sample size conditions (see Figure 12).

An interaction effect is apparent between the cross-level interaction and the treatment-control group ratio. Both X_1 and X_2, covariates with correlations with Z, show fewer significant γ_{10} as the ratio changes from 1:1 to 1:3 to 1:9, with this relationship steadily decreasing as the cross-level interaction increases. The percent of significant γ_{10} for X_3, which has an average null relationship with Z, shows a relationship with the ratio only when the cross-level interaction is at its strongest (0.3). These relationships are illustrated in Figure 13, below. The percentages of significant γ_{10} across simulation conditions when using between-cluster matching are presented in Figure 14, below.

Figure 11: Mean percent significant γ_{10} per treatment-control ratio and level-2 sample size.

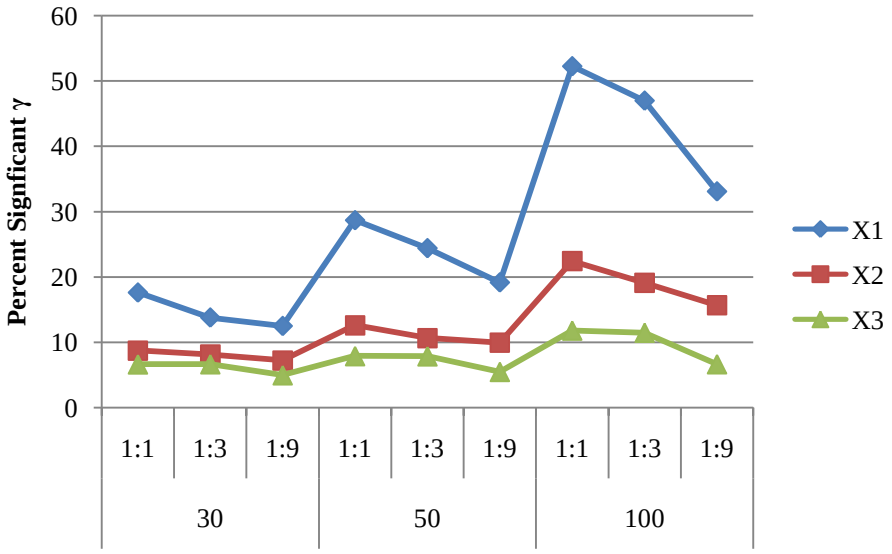


Figure 12: Mean percent significant γ_{10} per treatment-control ratio and level-1 sample size.

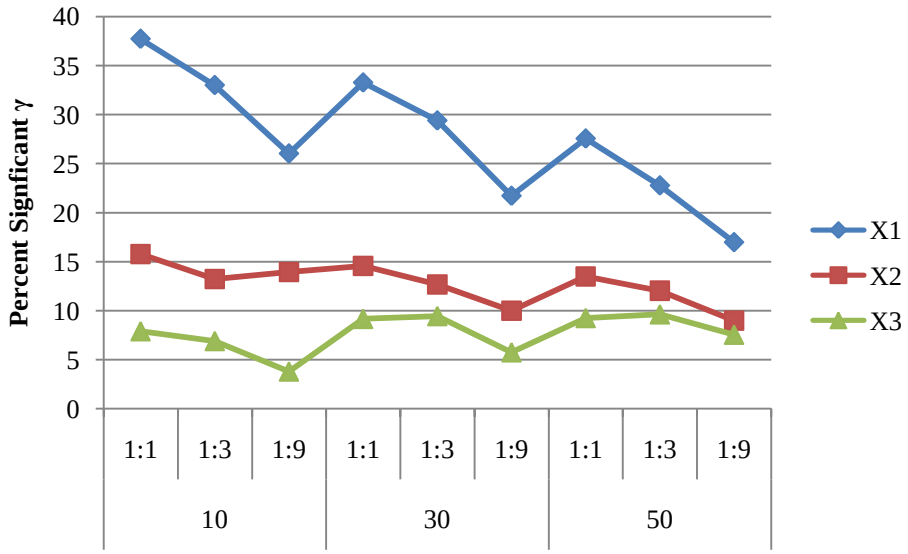


Figure 13: Mean percent significant γ_{10} per cross-level interaction and treatment-control group ratio.

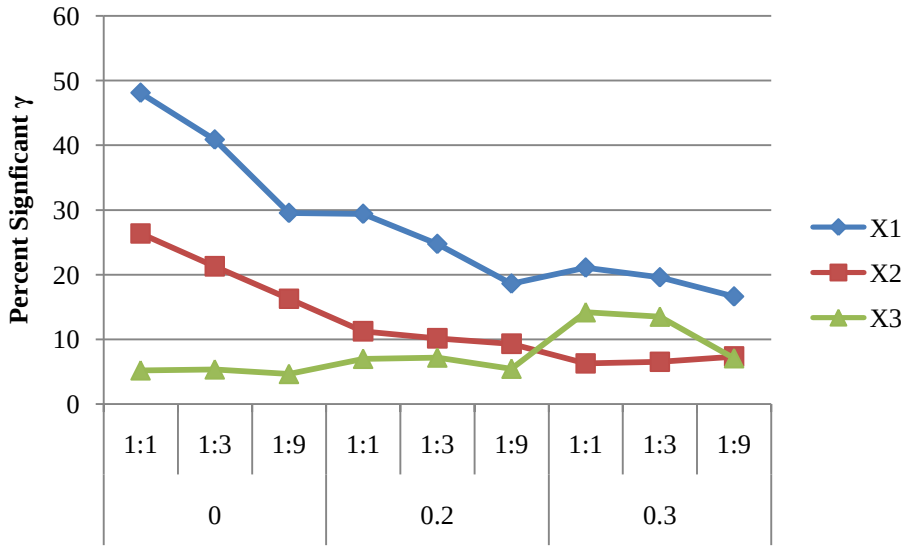
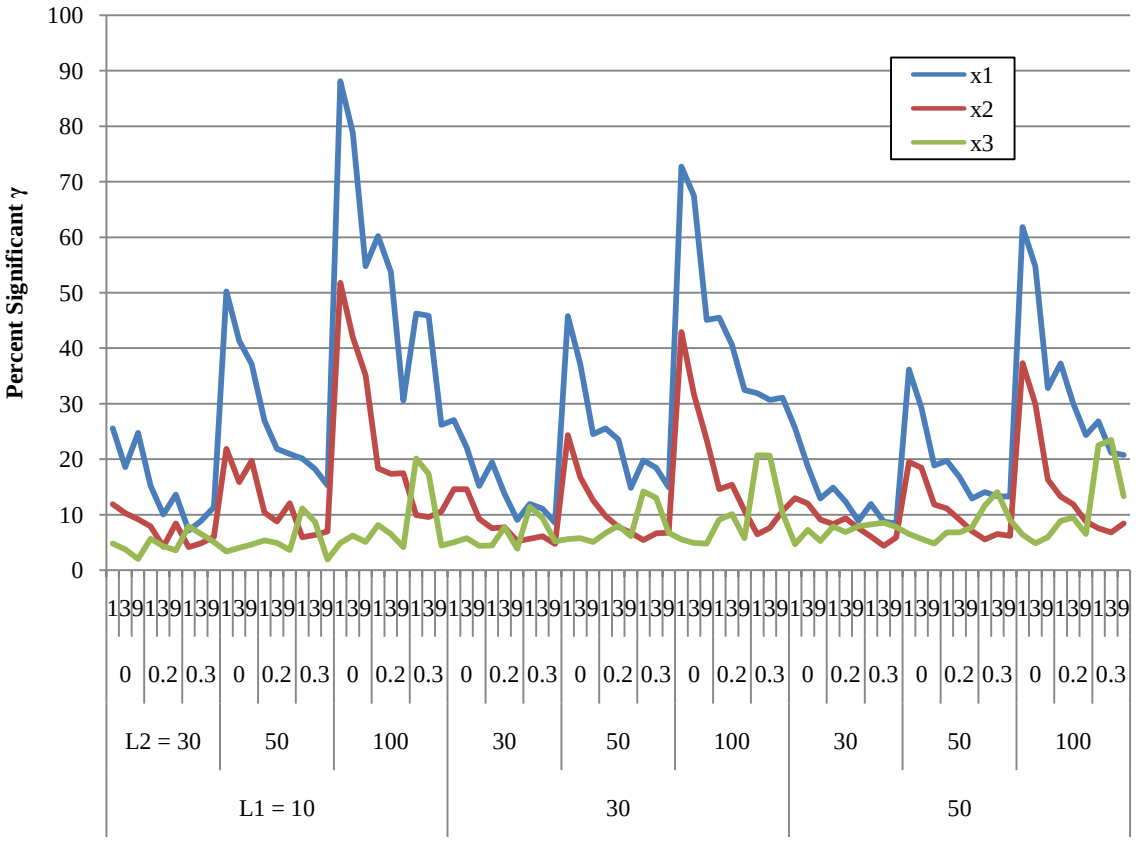


Figure 14: Percentage of Significant γ_{10} for X_1, X_2, and X_3 across Simulation Conditions for Between-Cluster Matching with a Propensity Score Estimated using Model 1



Variation in Balance Achievement within Clusters: Percent Significant τ_{11}

Sample Size at Level 1 and Level 2. When adjusting for selection bias by applying a propensity score estimated through a multilevel model to between-group matching, the percent significant τ_{11} is negatively related to both level-1 and level-2 samples: As sample size increases, the percent significant τ_{11} decreases. The relationship of the percent significant τ_{11} with the level-2 sample size is very small. The relationship of the percent significant τ_{11} with level-1 sample size is very small when increasing from 30 to 50 per group; the decrease in the percent significant τ_{11} is much larger when

increasing from 10 to 30 members per group. This pattern shows little variability across covariates X_1 through X_3.

Table 16:

Average Percentage of Significant τ_{11} Resultant from Between-Cluster Matching with a Propensity Score Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level 1 Sample Size			Level 1 Sample Size			Level 1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	12.1	4.2	3.6	12.7	3.5	3.2	12.8	3.8	2.8	6.5
50	12.0	3.4	4.4	11.6	2.7	3.4	10.3	2.2	3.3	5.9
100	8.7	2.9	5.3	8.4	1.9	4.4	8.7	1.6	4.0	5.1
Mean	10.9	3.5	4.4	10.9	2.7	3.7	10.6	2.5	3.4	5.9

Cross-Level Interaction. The strength of the cross-level interaction and the percent significant τ_{11} has a small but positive relationship, as evidenced in the final column of Table 17. The relationship of the cross-level interaction and percent significant τ_{11} is minimal to nonexistent in the five conditions with the largest sample sizes (all conditions when level-1 sample size is equal to 50, and the largest two sample sizes when the level-1 sample size is equal to 30). In these conditions, the percent significant τ_{11} fluctuates little more than 4% across conditions. When sample size at level-1 is smallest, the relationship between percent significant τ_{11} and the strength of the cross-level interaction is inconsistent.

Table 17:

Average Percentage of Significant τ_{11} for X_1 per Cross-Level Interaction Resultant from Between-Cluster Matching with a Propensity Score Estimated using Model 1

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	10.8	14.2	9.2	4.8	4.7	3.4	4.6	4.6	5.4	6.8
0.2	12.2	13.5	8.9	3.5	2.7	2.9	3.4	4.6	5.8	6.4
0.3	13.4	8.2	8.0	4.3	2.8	2.5	2.8	4.0	4.8	5.6

Treatment-Control Group Ratio. The relationship of the treatment-control group ratio and the percent significant τ_{11} is small and positive when the level-1 sample size is either 30 or 50. When sample size at level-1 is 10, however, the percent significant τ_{11} is elevated relative to other level-1 sample sizes. The percent is much higher than the other conditions when the treatment-control ratio is 1:9 (see Figure 15). The relationship of the treatment-control ratio and percent significant τ_{11} is consistent across level-2 sample sizes and across cross-level interactions, a finding that indicates no interaction effects (see Figure 16 and Figure 17). The percent significant τ_{11} is consistently largest when the ratio of treatment to controls is 1:9.

Figure 15: Mean Percent Significant τ_{11} per treatment-control ratio and level-1 sample size.

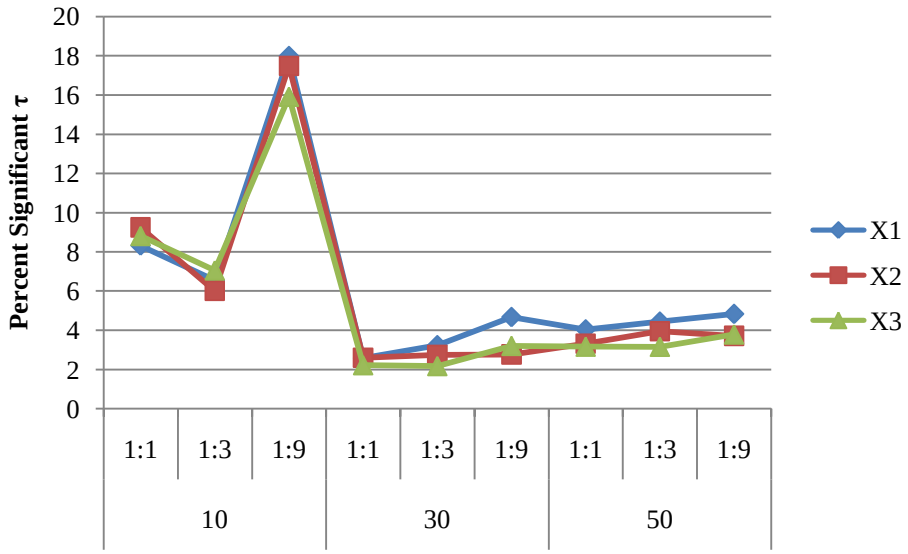


Figure 16: Mean Percent Significant τ_{11} per treatment-control ratio and level-2 sample size.

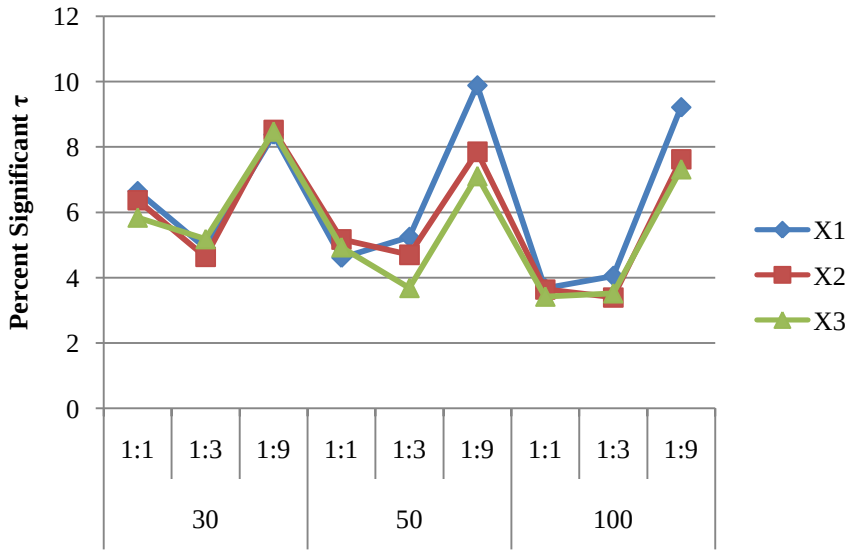
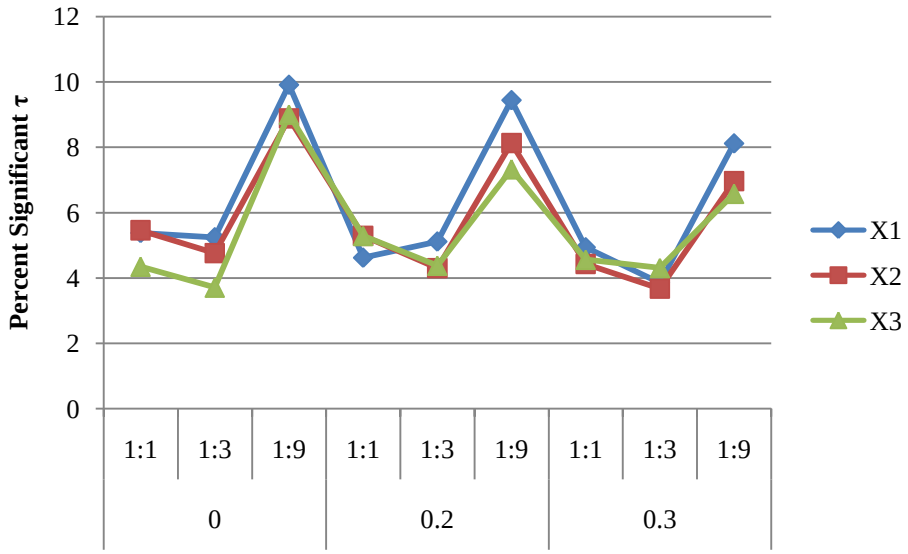


Figure 17: Mean Percent Significant τ_{11} per cross-level interaction and treatment-control group ratio.



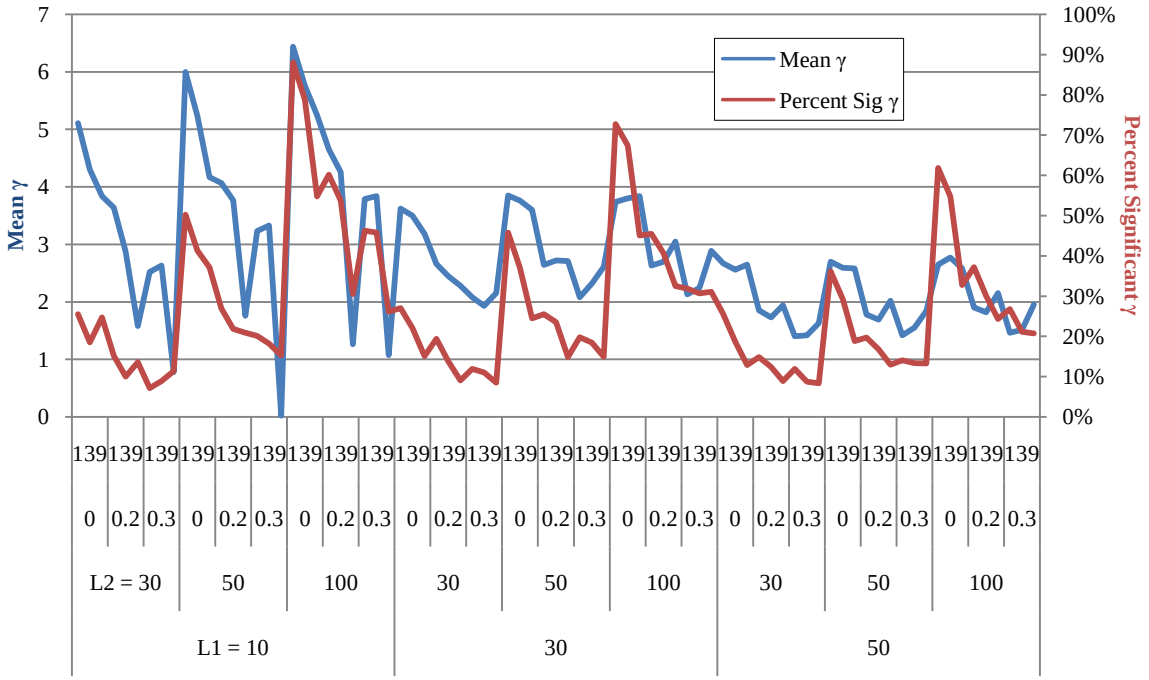
Percent Significant versus Mean Value of γ_{10} and τ_{11}

When using a propensity score estimated from a multilevel model with between-cluster matching, the pattern for the mean significant γ_{10} and the mean γ_{10} values are similar across conditions. This finding suggests that the proportion of significant γ_{10} is not influenced by sample-size. The performance of this propensity score and matching method in balancing covariates was best when the treatment-control group ratio was 1:9 (more control group members from which to match treatment group). This effect was most pronounced when the level-1 sample size was smallest. This pattern is opposite of that apparent when using within-cluster matching where balance was poorest in the 1:9 condition. The difference in results from these two adjustment methods is likely resultant from the fact that when matching between clusters, the entire sample is available from which to find close matches (e.g., not limited to matching only within the cluster). This finding suggests that propensity scores are estimated efficiently using a multilevel model,

even in the 1:9 condition, and the poorer performance of balancing when matching within-clusters is resultant from matches that are further apart on their propensity score than are available when matching across the sample as a whole.

The relationship between the cross-level interaction and the mean γ_{10} is also apparent in Figure 18, with greater cross-level interactions resulting in smaller values of γ_{10}. For an explanation of this relationship, consider the propensity score estimation model where the cross-level interaction has a mean value of 0 for each of the three level-1 covariates. In this case, including a covariate with no relationship with the treatment assignment mechanism (Z) would be equivalent to including an extraneous variable. In the estimated models, X_3 was this extraneous variable. When there was no cross-level interaction, X_3 was merely a nuisance variable, introducing noise into the propensity scores, which would result in poorer matches. When a cross-level interaction was present, however, X_3 was no longer merely a nuisance variable, but it provided information that resulted in more precise estimations of the propensity score and subsequently better matches.

Figure 18: Overlay of Mean γ_{10} to Percent Significant γ_{10} for Between-Cluster Matching with a Propensity Score Estimated using Model 1

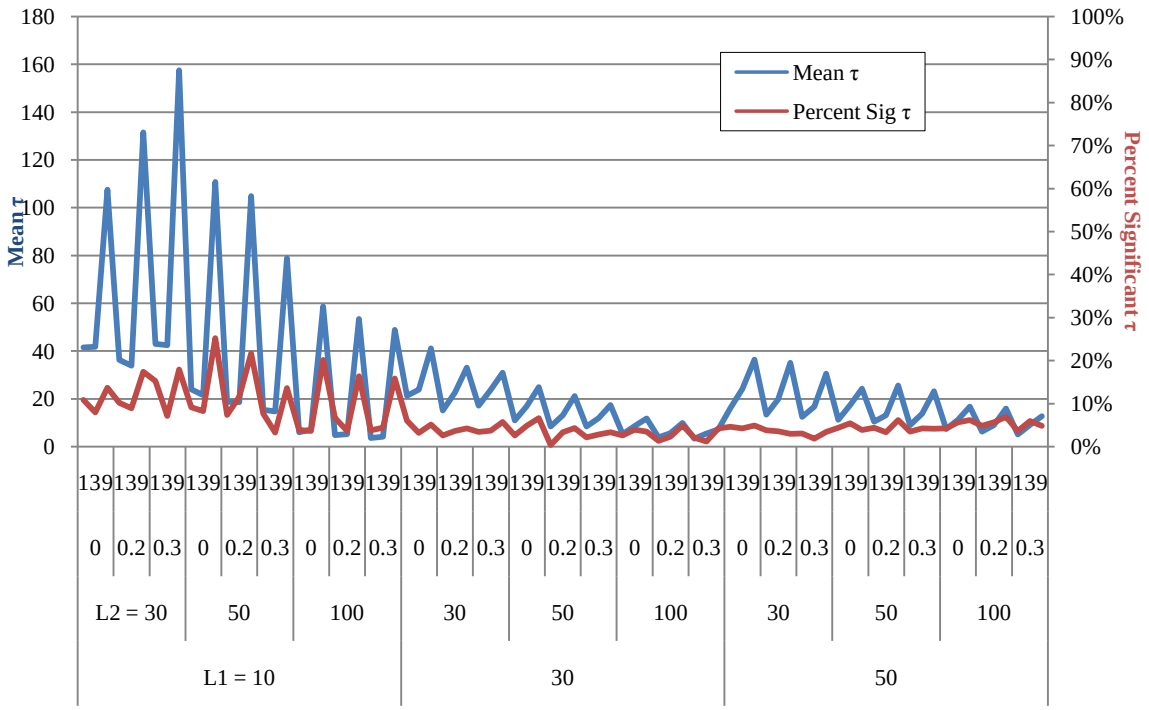


The mean significant τ_{11} and mean value of the τ_{11} per condition indicate that between-cluster matching has fairly consistent results within each cluster. As sample-size increases, variance in b_{1j} decreases. This finding is especially relevant when considering that matches are occurring across clusters. These small values of τ_{11} that are apparent under most conditions might be a result of the closest matches being within each cluster. Another possible explanation is that the propensity scores that are estimated through the multilevel model incorporate the cluster-level effects of the treatment assignment so that propensity scores of individuals from different clusters are equivalent.

Considering together the results for γ_{10} and τ_{11}, there is disagreement in the results related to the treatment-control group ratios in that balance across the sample as a whole is best in the 1:9 condition but the variance in balance is worst across clusters in the 1:9 condition. This finding might be a consequence of matching between clusters where

unequal ratios of treatment-control individuals exist within each cluster, or the extreme case where certain clusters have only treatment individuals and others have only control individuals. Estimations of overall balance may show adequate results while the balance within each cluster is not maintained. Once sample size overall reaches 3,000 (e.g., $n_{level\text{-}1}$ = 30 and $n_{level\text{-}2}$ = 100), however, the differences across treatment-control group ratios are minimal.

Figure 19: Overlay of Mean τ_{11} and Percent Significant τ_{11} for Between-Cluster Matching with a Propensity Score Estimated using Model 1



Balance Achievement in Predictor Covariates: The Standardized Mean Difference

Sample Size at Level 1 and Level 2. The relationship of level-1 sample size and level-2 sample size with SMD is negative when applying propensity scores estimated using multilevel modeling to between-cluster matching: as the number of individuals

increases, the SMD decreases. The relationship of SMD to level-2 sample size is only apparent in X_2 and X_3 where the correlation with the treatment assignment is smallest.

Table 18:

SMD for X_1 Resultant from Between-Cluster Matching with a Propensity Score Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	37.1	29.3	24.9	35.4	27.5	23.1	32.8	26.1	22.6	28.8
50	37.1	28.7	22.9	30.4	23.5	19.9	25.7	21.1	18.2	25.3
100	37.9	28.1	22.5	27.9	20.7	16.9	20.0	16.4	14.4	22.8
Mean	37.4	28.7	23.5	31.3	23.9	20.0	26.2	21.2	18.4	25.6

Cross-Level Interaction. The results of this study indicate that the strength of the cross-level interactions is negatively related to the SMD: as the cross-level interaction increases, the SMD decreases (see the final column in Table 19). This relationship is consistent across level-1 sample sizes and level-2 sample sizes. The average SMDs for covariate X_1 are presented in Table 19, below.

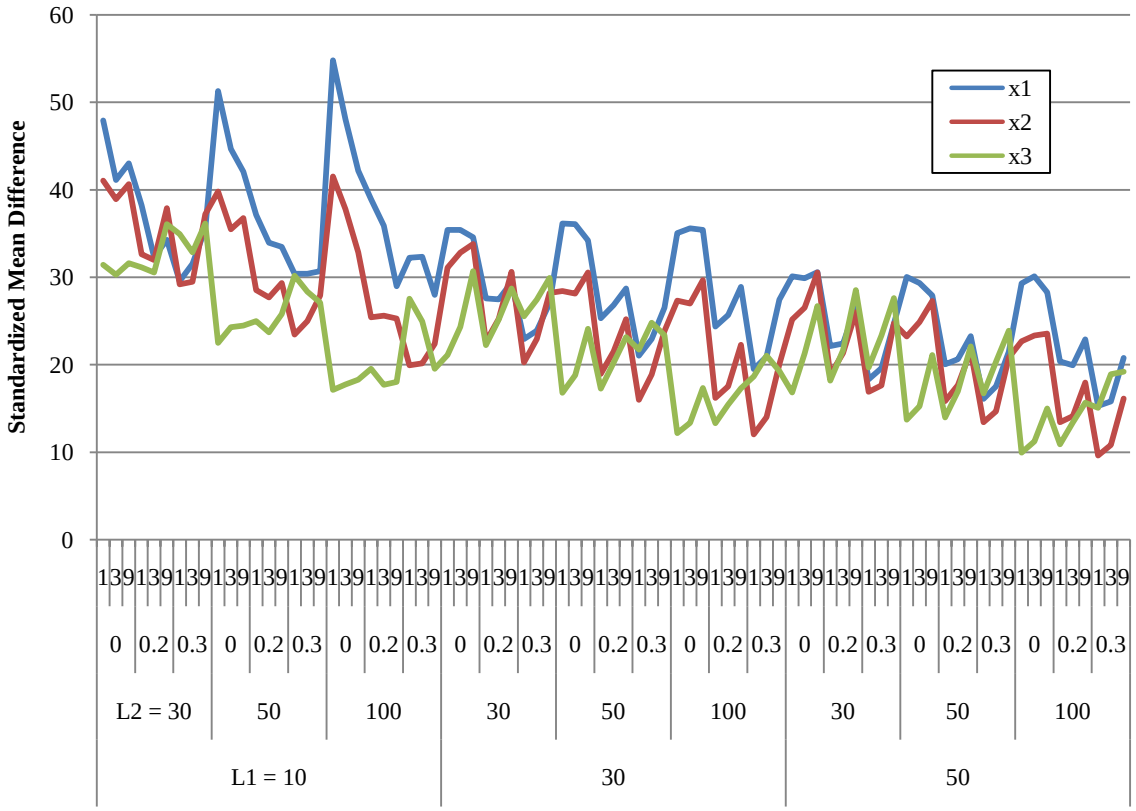
Table 19:

SMD for X_1 per Cross-Level Interaction Resultant from Between-Cluster Matching with a Propensity Score Estimated using Model 1

$\rho_{(wx)z}$	Level 1 Sample Size									Mean
	10			30			50			
	Level 2 Sample Size			Level 2 Sample Size			Level 2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	44.0	46.0	48.3	35.1	35.5	35.4	30.2	29.1	29.2	37.0
0.2	34.9	34.9	34.6	28.2	27.0	26.3	23.7	21.3	21.1	28.0
0.3	32.2	30.5	30.9	24.6	23.5	22.6	20.9	18.4	17.3	24.5

Treatment-Control Group Ratio. The relationship between the treatment-control group member ratios and the SMD showed an inconsistent pattern across sample-characteristics: When sample size at level-1 was 10, the SMD tended to be smallest in the 1:9 condition, whereas when sample sizes at level-1 were moderate and large, the SMD tended to be largest in the 1:9 condition. The exception to the pattern seen in the moderate and large level-1 sample size conditions was when the cross-level interaction was 0 in which case the SMD remained relatively unchanged across treatment-control group ratio conditions. Results for the SMD across conditions when using between-cluster matching are presented in Figure 20 below.

Figure 20: Standardized Mean Differences for X_1, X_2, and X_3 across Simulation Conditions using Between-Cluster Matching with a Propensity Score Estimated using Model 1

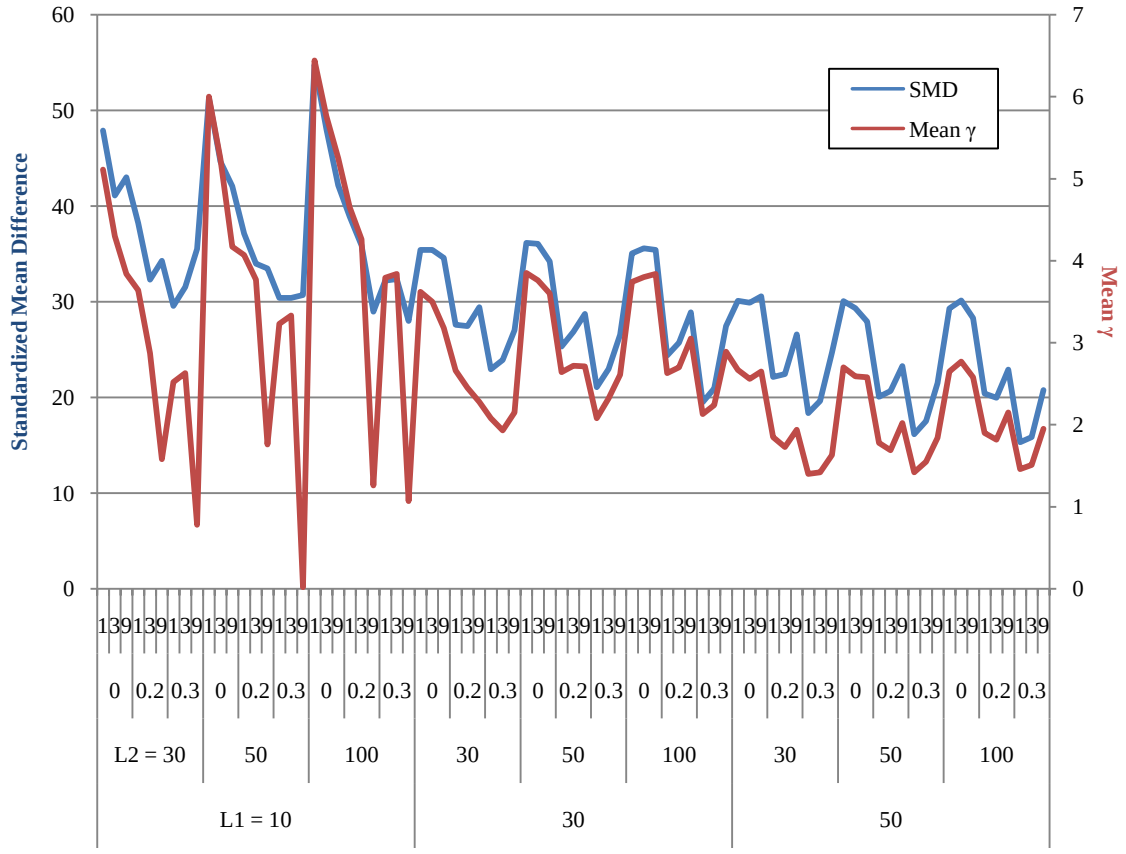


SMD versus Mean γ_{10}

As with within-cluster matching, the respective patterns of change for the SMD and mean γ_{10} are similar across between-cluster matching conditions once sample size at level-1 reaches 30. When sample size is small, the balance indicated by SMD shows less fluctuation across simulation conditions than is apparent in the mean values for γ_{10}. This pattern is similar to that found when using within-cluster matching as discussed previously. When level-1 sample size is small, the values for mean γ_{10} show particular sensitivity to the treatment-control group ratio where values for SMD appear more stable. This instability in the mean values of γ_{10} is likely related to challenges in estimating these

values when there are very few treatment group members rather than being reflective of the successful balance of the covariates.

Figure 21: Overlay of the SMD and Mean γ_{10} when using Between-Cluster Matching



Stratification into Quintiles

Balance Achievement in Predictor Covariates: Percent Significant γ_{10}

Sample Size at Level 1 and Level 2. The findings from this study indicate that the percentage of significant γ_{10} decreases as the sample-size at Level-1 increases. This pattern is evident in the final row of Table 20 for the mean γ_{10} for X_1. As the covariates' correlations with treatment assignment decreases, however, the relationship between level-1 sample size and percentage of significant γ_{10} becomes non-linear. The relationship

of the level-2 sample size with the percent of significant γ_{10} across simulation replications was positive: as the level-2 sample size increased, the percentage of significant γ_{10} increased. This relationship was more pronounced when the level-1 sample size was small (see Table 20).

Table 20:

Average Percentage of Significant γ_{10} Resultant from Quintile Stratification with a Propensity Score Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	14.9	16.0	12.5	5.6	8.3	7.8	3.1	5.8	6.6	9.0
50	30.2	24.1	13.8	10.4	10.0	8.6	3.3	6.4	6.8	12.6
100	59.1	36.5	17.4	22.4	13.9	8.8	6.3	8.7	8.4	20.2
Mean	34.7	25.5	14.6	12.8	10.7	8.4	4.2	7.0	7.3	13.9

Cross-Level Interaction. The results of this study indicate that the strength of the cross-level interactions are negatively related to the percentage of significant γ_{10} when propensity scores estimated using a multilevel model are applied using quintile stratification: as the cross-level interaction increases, the percent of significant γ_{10} decreases. This relationship was most pronounced among the γ_{10} of X_1, which had the largest correlation with the treatment assignment, followed by X_2, and little relationship indicated for X_3. The percentage of significant γ_{10} per sample characteristics for X_1 are presented in Table 21 below. The final column of this table illustrates the decrease in the

percentage of significant γ_{10} from the simulation replications as the cross-level interaction increases. The pattern was stable across conditions.

Table 21:

Average Percentage of Significant γ_{10} for X_1 per Cross-Level Interaction Resultant from Quintile Stratification with a Propensity Score Estimated using Model 1

$\rho_{(WX)Z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	20.9	45.0	75.1	21.6	33.3	47.8	14.7	17.5	22.5	33.1
0.2	13.8	25.7	54.7	14.7	20.4	32.2	11.7	12.0	15.4	22.3
0.3	10.0	19.9	47.5	11.7	18.5	29.5	11.1	11.8	14.3	19.4

Treatment-Control Group Ratio. The ratio of treatment group members and control group members has a bearing upon the percentage of significant γ_{10} under specific conditions. The finding for quintile stratification is similar to that found in between-cluster matching, and opposite that found in within-cluster matching: The magnitude of the differences between the number of treatment and control group members within each cluster is negatively related to the percentage of significant γ_{10}. In other words, as the number of control individuals becomes greater relative to the number of treatment individuals, the percentage of significant γ_{10} decreases. Additionally, the relationship of the treatment-control group ratio and the percentage of significant γ_{10} shows an interaction with level-2 sample size, becoming more pronounced as level-2 sample size increases, and with level-1 sample size, becoming less pronounced as level-1 sample size

increases. The interaction of the treatment-control group ratio and cross-level interaction and the percent significant γ_{10} is nearly identical to those apparent in the between-group matching. These relationships are presented in Figure 22, Figure 23, and Figure 24, below. The percentage of significant γ_{10} associated with each condition when using stratification is presented in Figure 25 below.

Figure 22: Mean percent significant γ_{10} per treatment-control ratio and level-2 sample size.

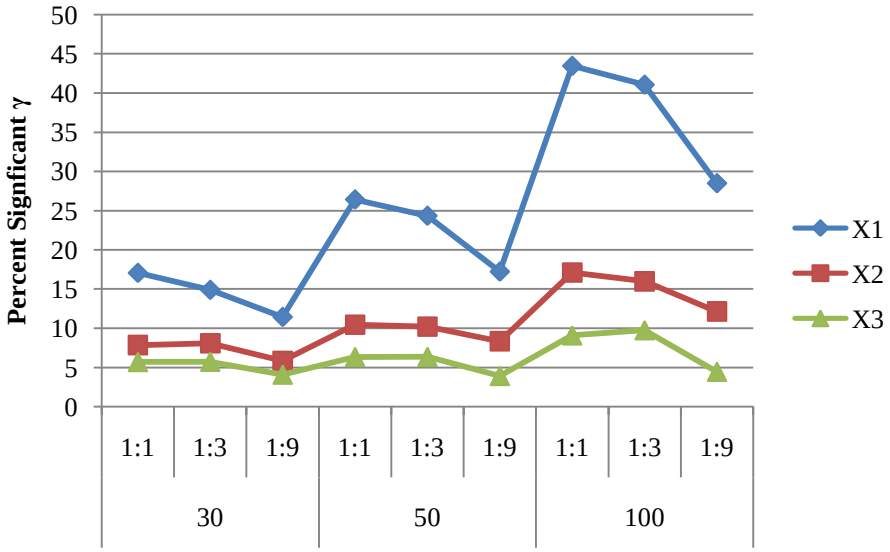


Figure 23: Mean percent significant γ_{10} per treatment-control ratio and level-1 sample size.

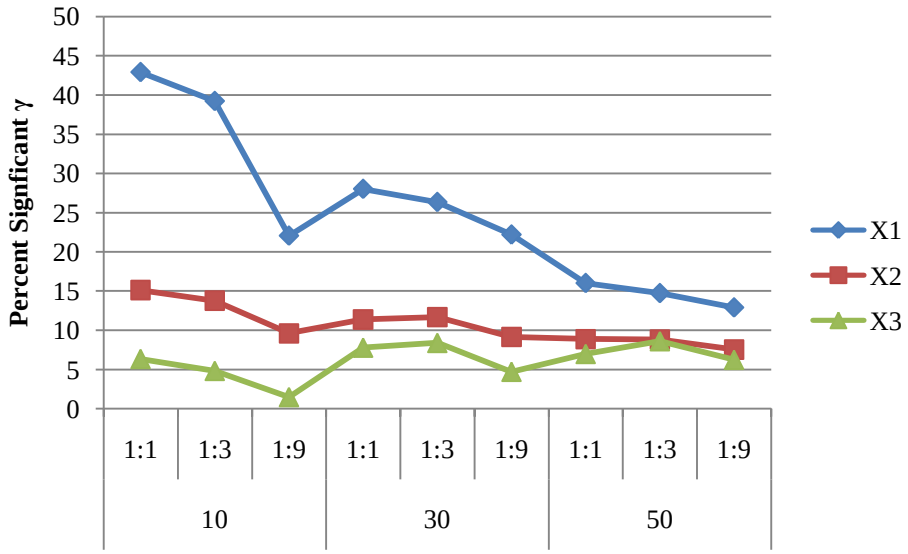


Figure 24: Mean percent significant γ_{10} per cross-level interaction and treatment-control ratio

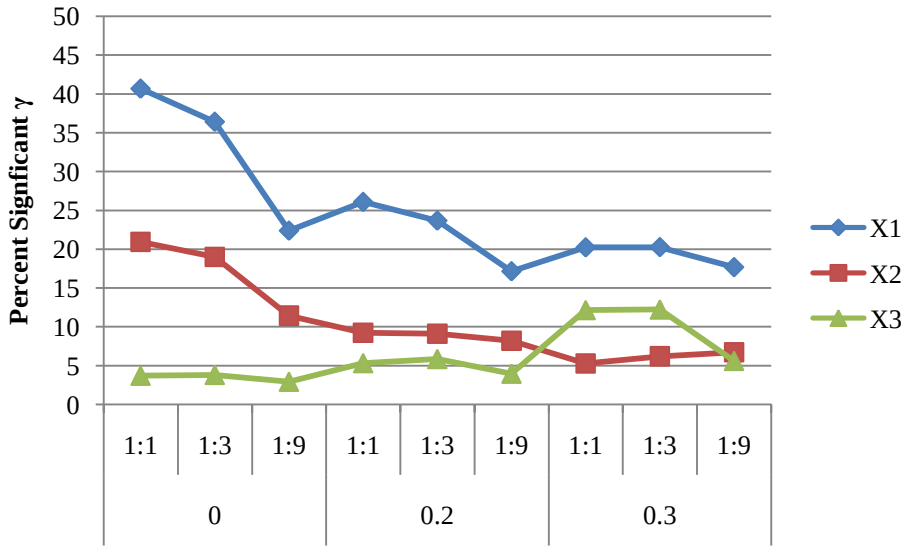
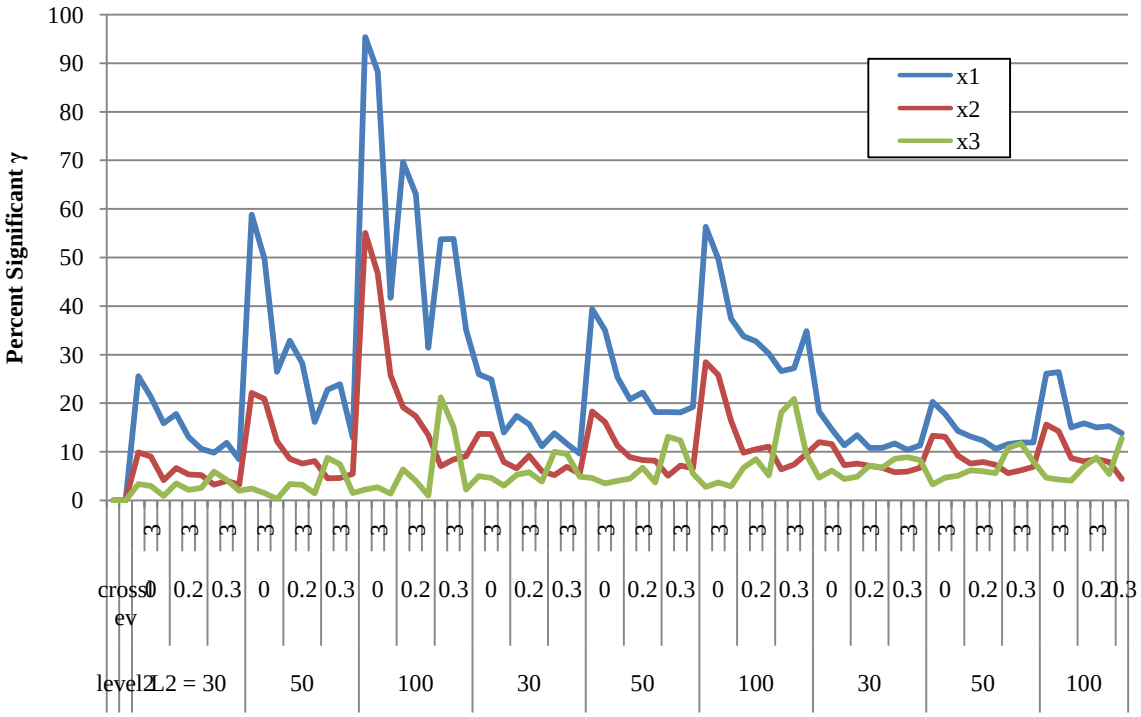


Figure 25: Percentage of Significant γ_{10} for X_1, X_2, and X_3 across Simulation Conditions for Quintile Stratification with a Propensity Score Estimated using Model 1



Variation in Balance Achievement within Clusters: Percent Significant τ_{11}

Sample Size at Level 1 and Level 2. The relationship between percent significant τ_{11} and level-2 sample size is positive: as the level-2 sample size increases, the percent significant τ_{11} increases. The relationship between the level-1 sample size and percent significant τ_{11} is non-linear: when level-1 sample size is moderate (n=30), the percent significant τ_{11} is consistently smallest compared to level-1 sample sizes of 10 and 50. This pattern is consistent across covariates X_1 through X_3 (see Table 22).

Table 22:

Average Percentage of Significant τ_{11} Resultant from Quintile Stratification with a Propensity Score Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	14.1	2.7	8.5	10.2	2.3	6.9	9.8	2.0	6.8	7.0
50	20.0	8.7	20.4	16.4	5.6	18.8	15.2	6.2	19.1	14.5
100	29.2	25.1	49.8	23.2	20.2	43.7	24.1	20.0	43.5	31.0
Mean	21.1	12.1	26.2	16.6	9.4	23.2	16.4	9.4	23.1	17.5

*Cross-Level Interaction.*The relationship of the cross-level interaction and percent significant τ_{11} is small under most sample size conditions. When sample size at level-1 is moderate (n=30), the strength of the cross-level interaction has no relationship with the percent significant τ_{11}. A consistent pattern is only evident when the level-1 sample size is smallest and level-2 sample size is moderate and large. Under these conditions, as the strength of the cross-level interaction increases, the percent significant τ_{11} decreases (see Table 23).

Table 23:

Average Percentage of Significant τ_{11} for X_1 per Cross-Level Interaction Resultant from Quintile Stratification with a Propensity Score Estimated using Model 1

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	15.9	25.5	37.8	2.4	8.4	24.0	7.4	18.4	46.1	20.7
0.2	11.8	18.5	28.5	2.7	8.8	25.8	8.3	20.3	49.8	19.4
0.3	14.6	16.0	21.2	2.9	8.8	25.4	9.8	22.5	53.5	19.4

Treatment-Control Group Ratio The relationship of treatment-control group ratio and percent significant τ_{11} varies across level-1 sample sizes when stratifying on the propensity score. When sample size is smallest (n=10), the relationship between percent significant τ_{11} and treatment-control group ratio is positive, whereas when the level-1 sample size is either 30 or 50, the relationship is negative: as the difference between treatment group and control group members increases, the percent significant τ_{11} decreases (see Figure 26). This pattern is opposite that indicated for within-group matching.

The findings of this study indicate that an interaction effect does not exist with level-2 sample size and the relationship between treatment-control group ratio and percent significant τ_{11} (see Figure 27). The mean percent significant τ_{11} is consistently largest when the ratio is 1:9; and the mean percent significant τ_{11} is smallest when the ratio is 1:3.

A small interaction effect is apparent between the cross-level interaction and the treatment-control group ratio on the mean percent significant τ_{11}. As the cross-level interaction increases, the relationship between the treatment-control group ratio and the mean percent significant τ_{11} becomes smaller, as illustrated in Figure 28.

Figure 26: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-1 sample size.

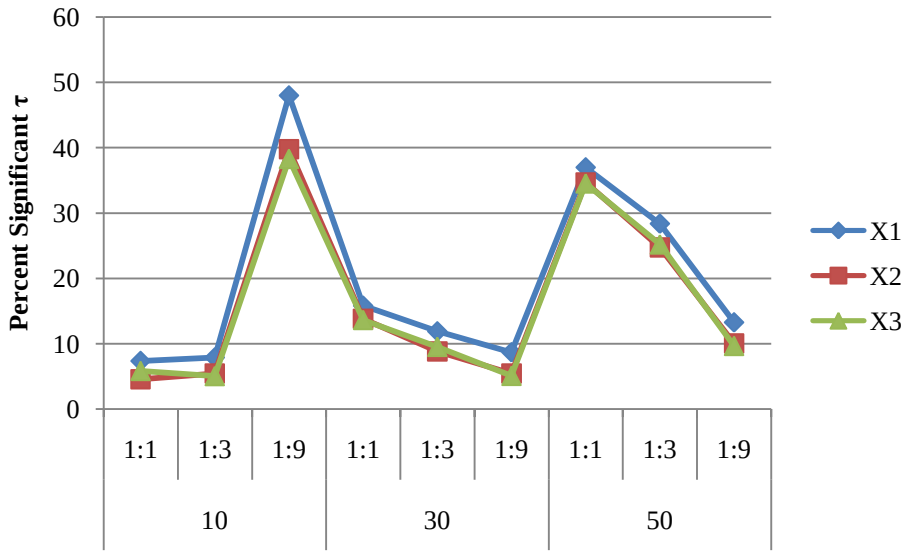


Figure 27: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-2 sample size.

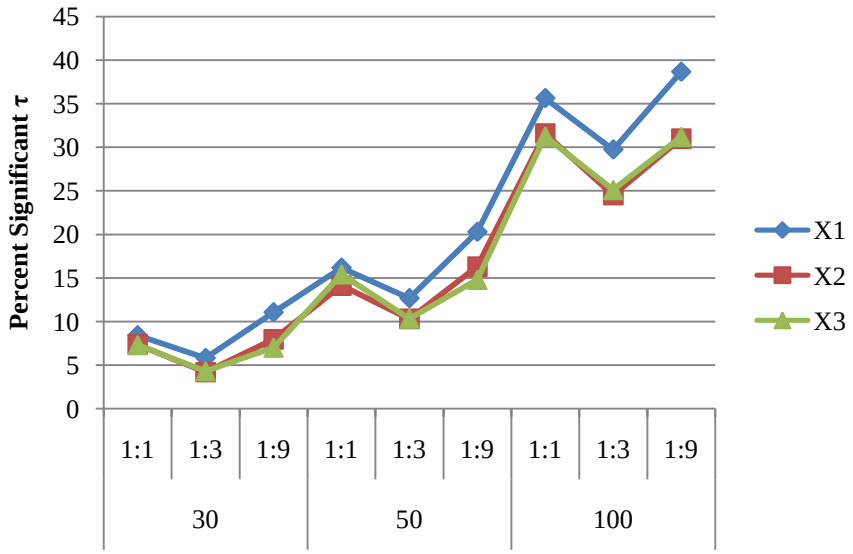
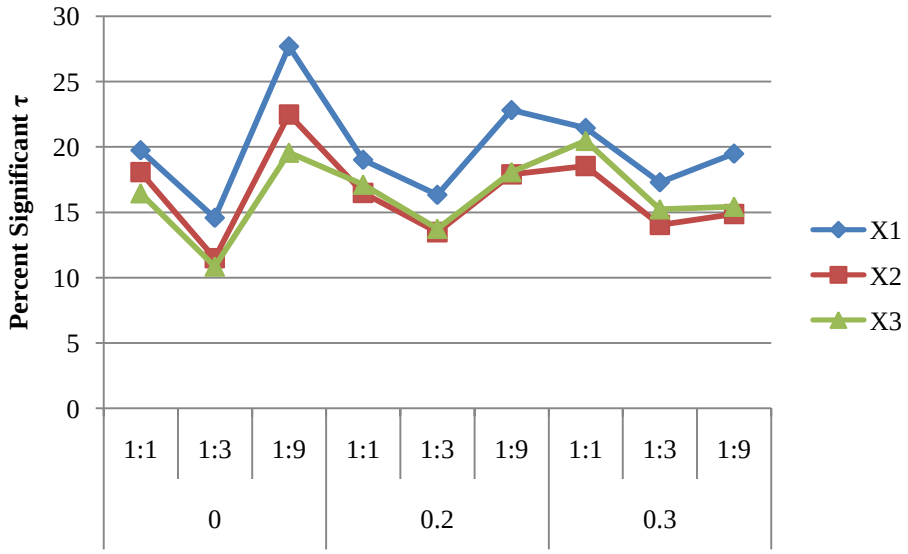


Figure 28: Mean Percentage of Significant τ_{11} per cross-level interaction and treatment-control group ratio.

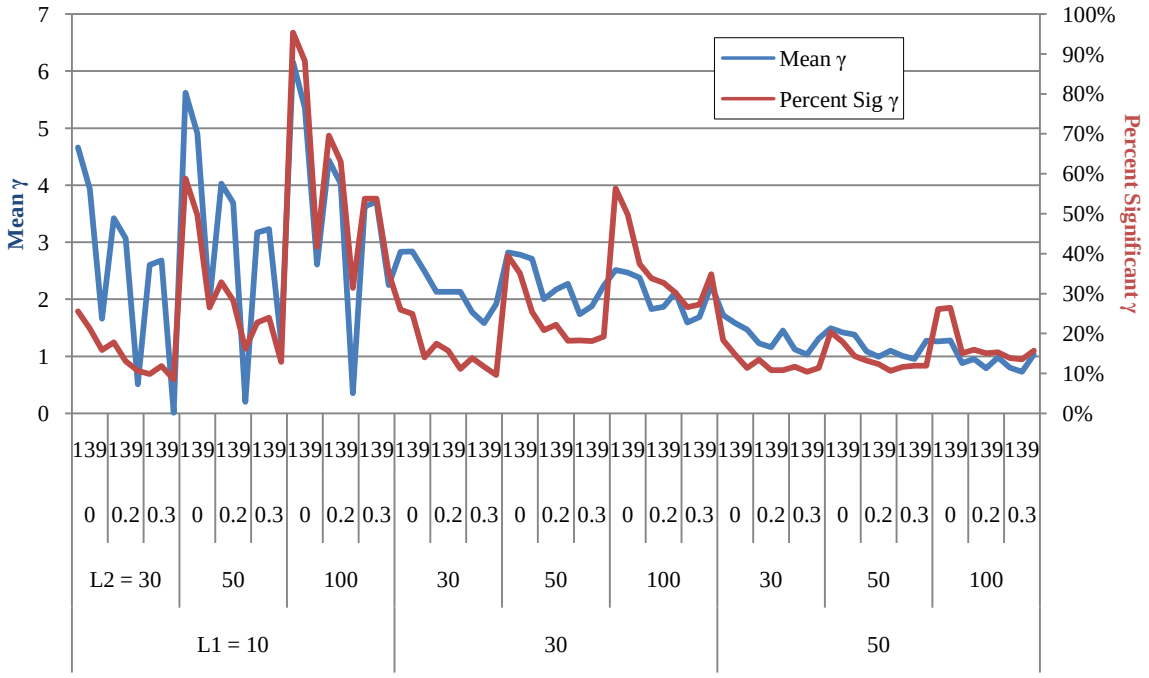


Percent Significant versus Mean Value of γ_{10} and τ_{11}

The pattern for mean γ_{10} is similar to that of percent significant γ_{10} when applying a propensity score estimated from a multilevel model using quintile stratification. Under these conditions, the performance in balancing covariates shows great fluctuation across conditions when sample size is small and more stable results across conditions once sample size at level 1 reaches 30. Even when sample size is small at level 1, however, this method shows favorable results when the number of control group members is large compared to that of the treatment group sample size (1:9 ratio). This finding is congruent to that found with between-cluster matching and opposite that found with within-cluster matching. This finding suggests that having a greater number of control individuals from which to compare treatment group individuals results in better covariate balance across the overall sample even when the sample is rather small.

These favorable results are further improved when cross-level interactions are larger ($\rho_{(WX)Z}$ =0.3). This finding is also congruent with that of between-cluster matching, but this relationship is not apparent with within-cluster matching. This suggests that incorporating a cross-level interaction term into the propensity score estimation model is most beneficial when matches are to be attained across clusters. When matching within clusters, including the cross-level interactions in the estimation equation does not have as great of a bearing upon the overall balance.

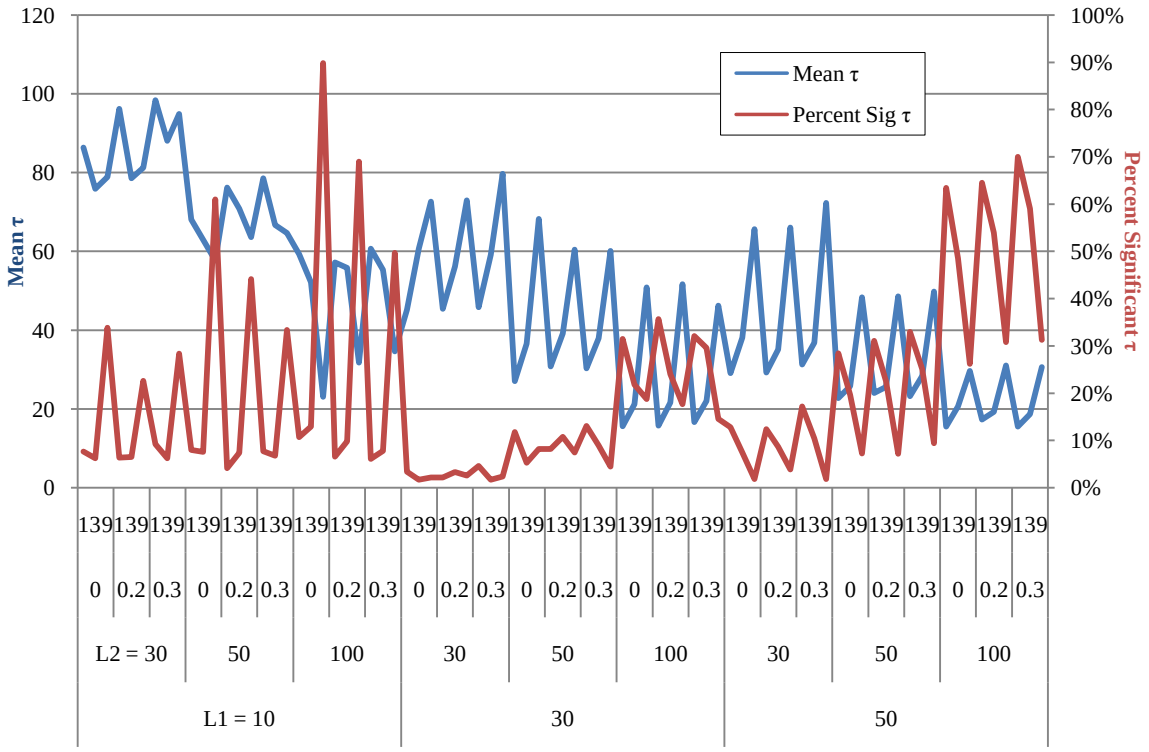
Figure 29: Overlay of Mean γ_{10} and Percent Significant γ_{10} for Quintile Stratification with a Propensity Score Estimated using Model 1



When considering the variance in the values of γ_{10} across clusters, as evidenced by the values of τ_{11}, opposite patterns are apparent in the mean values of τ_{11} versus the percentage of significant τ_{11}. This pattern is illustrated in the 1:9 treatment-control ratio conditions which almost always show the largest mean value of τ_{11} while showing the smallest percentage of significant τ_{11}. Because the variance across clusters is smallest under the same conditions where it is also most likely to be found significant, the likely explanation is related to sample size. If overall sample sizes are larger in the 1:1 ratio condition than under other ratio conditions, the increased percent of significant τ_{11} is reflective of the power to detect differences in the slopes across clusters rather than in the magnitude of the variances. Such an explanation is supported by the fact that, within each level-1 sample size condition, the percent significant τ_{11} is largest when overall sample-

size is largest while the mean values for τ_{11} are smallest when overall sample-size is largest (see Figure 39 below).

Figure 30: Overlay of Mean τ_{11} and Percent Significant τ_{11} for Quintile Stratification with a Propensity Score Estimated using Model 1



Balance Achievement in Predictor Covariates: The Standardized Mean Difference

Sample Size at Level 1 and Level 2. The relationship of level-1 sample size and level-2 sample size with SMD is negative when applying propensity scores estimated using multilevel modeling through quintile stratification: as the number of individuals increases, the SMD decreases. This relationship is consistent across sample-size conditions. Considering an SMD of 10 as indicating imbalance (D'Agostino & Rubin, 2000; Rosenbaum & Rubin, 1985), imbalance remained in all covariates across level-1 and level-2 sample characteristics (see Table 24).

Table 24:

SMD Resultant from Quintile Stratification with a Propensity Score Estimated using Model 1

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	55.3	43.6	38.4	50.2	38.8	34.5	47.8	37.5	33.7	42.2
50	49.4	39.3	34.4	41.5	33.6	30.6	38.6	31.8	29.5	36.5
100	46.3	35.8	30.9	34.6	28.6	26.3	29.5	26.7	25.1	31.5
Mean	50.3	39.5	34.5	42.1	33.7	30.4	38.6	32.0	29.5	36.7

Cross-Level Interaction. The results of this study indicate that the strength of the cross-level interactions is negatively related to the SMD: as the cross-level interactions increase, the SMD becomes smaller. This relationship is consistent across level-1 sample sizes and level-2 sample sizes. The average SMDs for X_1 are presented in Table 25, below.

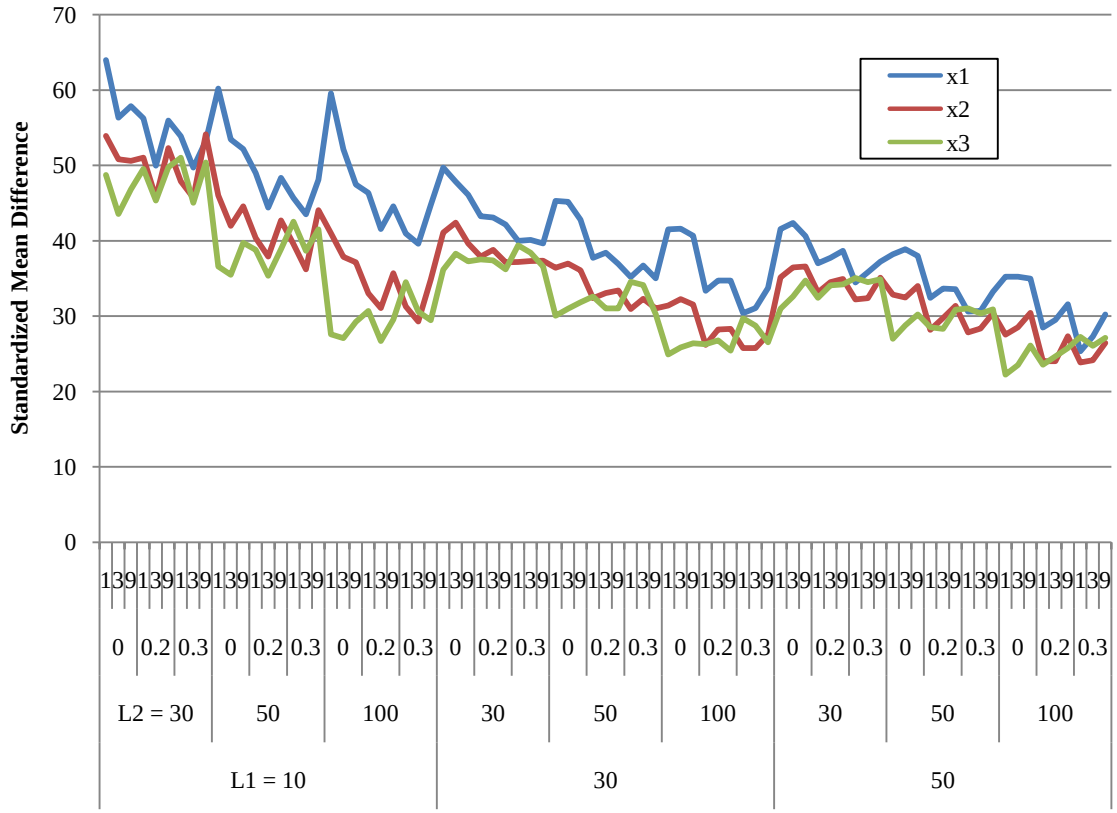
Table 25:

SMD for X_1 per Cross-Level Interaction Resultant from Quintile Stratification with a Propensity Score Estimated using Model 1

$\rho_{(WX)Z}$	Level 1 Sample Size									Mean
	10			30			50			
	Level 2 Sample Size			Level 2 Sample Size			Level 2 Sample Size			
	30	50	100	30	50	100	30	50	100	
0.0	59.4	55.3	53.1	47.9	44.4	41.3	41.5	38.4	35.2	46.3
0.2	54.1	47.3	44.2	42.9	37.7	34.3	37.8	33.2	29.9	40.1
0.3	52.3	45.8	41.8	39.9	35.7	31.8	35.8	31.5	27.6	38.0

Treatment-Control Group Ratio. The relationship between the treatment-control group member ratios showed a consistent pattern across sample-characteristics: larger differences between the numbers of treatment group versus control group members are positively related to SMD. In other words, as the number of control group members becomes larger relative to the number of treatment group members, the SMD decreases. Additionally, findings indicate that sample characteristics in which level-1 sample size is small and level-2 sample size is large, the relationship of the treatment-control group ratio and percentage of SMD is most pronounced. Values for the SMD across sample conditions for quintile stratification are presented in Figure 31 below.

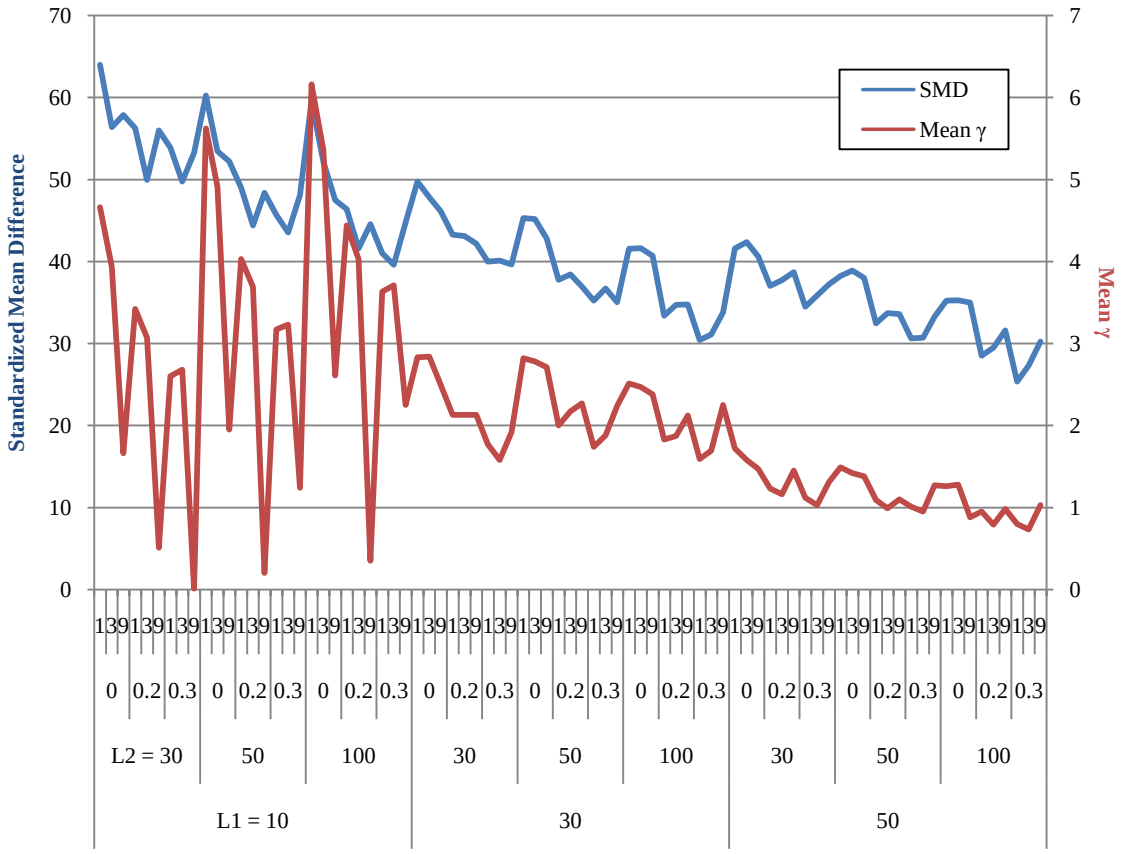
Figure 31: Standardized Mean Differences for X_1, X_2, and X_3 across Simulation Conditions using Quintile Stratification with a Propensity Score Estimated using Model 1



SMD versus Mean γ_{10}

When comparing the SMD to the mean γ_{10}, the same pattern is apparent as with between-cluster matching and within-cluster matching. With quintile stratification, the values are parallel once sample size at level-1 has reached 30. The SMD shows less fluctuation across conditions when sample size is small. The relationship of covariate balance with the cross-level interaction and with the treatment-control group ratio is apparent in Figure 32 below.

Figure 32: Overlay of the SMD and Mean γ_{10} when using Quintile Stratification



Comparison of the Performance of the Adjustment Methods

Retaining Treatment Group Members. Another approach to measuring the performance of propensity scores that are estimated using a multilevel model in attaining balance in covariates is to explore the retained sample size. This is not used as a measure of success, but rather as a measure of the consequences of propensity score adjustment. A small percentage of retained treatment group members compared to the initial sample may be indicative of the sample having few appropriate matches and may also have adverse effects on external validity if the retained treatment group members are very different from the treatment group in the sample as a whole. The propensity score

adjustment method that retained the greatest percentage of the initial treatment group was quintile stratification (46.3%), followed by between-cluster matching (average percent retained 22.9%), and within-cluster matching (average percent retained 8.8%).

When using quintile stratification, a positive relationship existed between the level-1 sample size and the percent treatment group retained and level-2 sample size and percent treatment group retained. When within-cluster matching was used, the relationship of percent treatment group retained and level-1 sample size was positive. The relationship with level-2 sample size was only relevant when overall sample size was small, in which case the relationship was also positive. When considering Figure 33 below, this benefit to percent retained that was resultant from increasing the number of clusters was only apparent when the treatment-control group ratio was 1:9. This suggests that propensity scores were more similar within clusters when fewer treatment group members were present per cluster. When coupled with the fact that overall balance was poorest for within-cluster matching in the 1:9 condition, the findings suggest that the propensity scores were less able to adequately estimate the treatment assignment mechanism for the treatment group members. Subsequently, their propensity scores were more similar to the control group members, although their covariates were not. The strength of the cross-level interaction had no relationship with the percent treatment group retained.

When between-cluster matching was used, the relationship of percent treatment group retained with level-1 sample size was negative and with level-2 sample size was positive. The relationship between retained percentage and treatment-control ratio was also most pronounced when between-cluster matching was used, with a much larger

percentage being retained in the 1:9 condition than in the 1:1 condition. The 1:9 condition with between-cluster matching is also the only time when the cross-level interaction shows a relationship with percent treatment group retained; in this case, the relationship is positive. Considering that the greatest overall balance was achieved in the 1:9 condition, the results suggest that between-cluster matching greatly benefits from having a relatively larger pool of control group members with whom to match treatment group members. As sample size increased, the relationship of percent treatment group retained and the treatment-control group ratio became less pronounced.

Larger sample sizes were a clear benefit to quintile stratification for retaining treatment group members. A possible explanation of this benefit is likely related to the source of sample loss when using stratification: lack of overlap in propensity scores between the treatment group and the control group. A likely conclusion, therefore, is that propensity scores estimated from larger samples resulted in less variability than propensity scores in smaller samples. A related explanation is that larger samples resulted in propensity scores for control group members and treatment group members that were more similar than when smaller samples were used. As with between-cluster matching, larger sample sizes showed a decrease in the relationship of percent treatment group retained and the treatment-control group ratio. Also like between-cluster matching, the cross-level interaction was positively related to the percent treatment group retained.

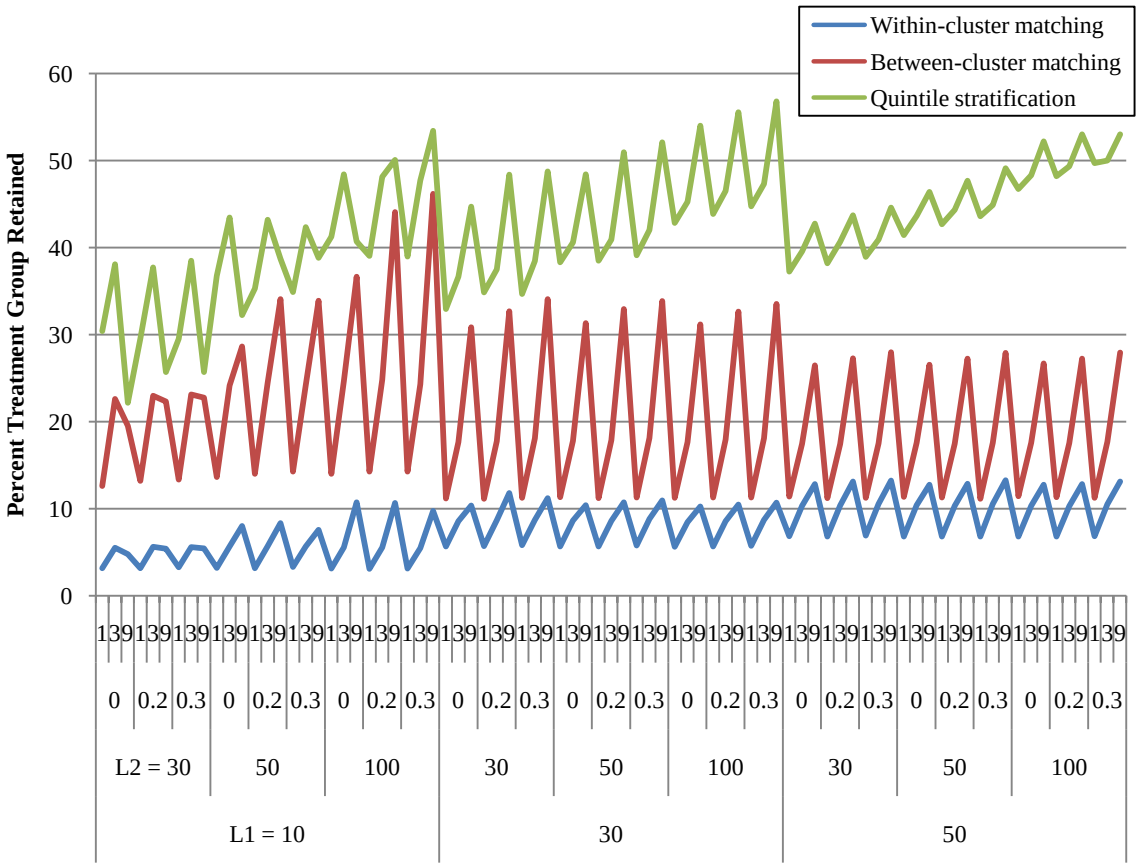
Table 26:

Percent of the Initial Treatment Group Retained per Propensity Score Adjustment Method and Sample Size Conditions

Level-2 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification			Mean
	Level 1 Sample Size			Level 1 Sample Size			Level 1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	4.7	8.5	10.1	19.2	20.5	18.7	30.8	39.7	40.7	21.4
50	5.7	8.4	10.1	23.5	20.7	18.7	38.4	43.4	44.9	23.7
100	6.4	8.3	10.1	27.0	20.6	18.8	45.3	48.5	50.1	26.1
Mean	5.6	8.4	10.1	23.2	20.6	18.7	38.2	43.9	45.2	23.8

Several relationships are readily apparent upon first glance at Figure *3*333. First, stratification retained more treatment group members than did either matching technique, whereas between-cluster matching retained more treatment group members than did within-cluster matching. The relationship of the treatment-control ratio is also clearly illustrated, with greater treatment group members retained when there were larger numbers of control group members with whom to match; however, the relationship between treatment-control group ratio and treatment group retention decreased as the level-1 sample size increased. The relationship of the cross-level interaction and treatment group sample size is greatest in stratification, and is large across propensity score adjustment methods when sample-size at level-1 is small. Similarly, the relationship of level-2 sample size and treatment group sample size is apparent in stratification, and across adjustment methods when level-1 sample size is small.

Figure 33: Percent of Treatment Group Retained across Simulation Conditions per Propensity Score Adjustment Method using a Propensity Score Estimated using Model 1



Cluster Retention Rate. The performance of the propensity score adjustment methods can also be described through its retention of the initial number of clusters. These percentages for each condition are presented in Figure 34 below. Quintile stratification resulted in the greatest percentage of retained clusters, followed by between-group matching, and within-group matching. The relationship between percentage retained and level-1 sample size is readily apparent: as level-1 sample size increases, the percentage of the retained clusters increases. For between-group matching and stratification methods, this relationship was greatly reduced in strength once the level-1 sample size reached 30.

The relationship with the level-2 sample size and the percentage of groups retained can be observed in between-group matching and stratification when level-1 sample size is small. In the remaining conditions, the level-2 sample size has a decreased relationship with the percent retained. The relationship of the percentage retained and the cross-level interactions was positive, though small: as the cross-level interactions increased, the percentage of the clusters retained also increased.

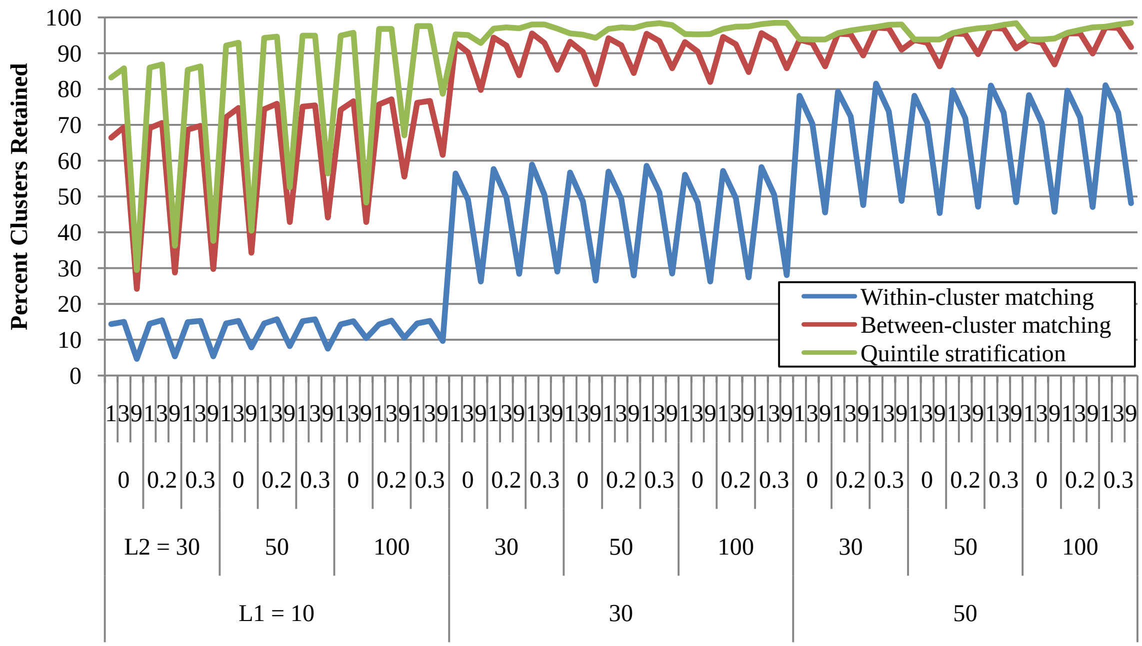
The treatment-control group member ratio showed different relationships with the percentage retained across simulation conditions. When level-1 sample size was small, the treatment-control ratio showed a similar pattern across propensity score adjustment methods: a slight increase in the retention rate is indicated as the ratio changed from 1:1 to 1:3, whereas a substantial decrease in sample size retained is related to the 1:9 treatment-control ratio condition. For between-group matching and within-group matching, the 1:9 treatment-control ratio condition is consistently associated with the smallest percentage retained. Once the level-1 sample size reaches 30 for the two matching methods, the 1:1 ratio is associated with the greatest percent retention and the 1:3 ratio is associated with the moderate retention rate. For quintile stratification, once the level-1 sample size reaches 30, the cluster retention rate is consistently above 90%. The retention rate for stratification shows a slight positive relationship with both the cross-level interaction and the treatment-control ratio.

Table 27:

Percent of the Initial Number of Clusters Retained per Propensity Score Adjustment Method and Sample Size Conditions

Level-2 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
30	11.7	45.1	66.4	55.2	89.7	93.1	68.6	96.4	96.0	69.1
50	12.8	44.9	66.2	63.2	90.0	93.2	79.3	96.7	96.0	71.4
100	13.3	44.6	66.2	68.5	90.2	93.4	86.0	97.0	96.2	72.8
Mean	12.6	44.9	66.3	62.3	90.0	93.2	77.9	96.7	96.1	71.1

Figure 34: Percent of the initial clusters retained per Simulation Condition and Propensity Score Adjustment Method with a Propensity Score Estimated using Model 1



A pattern in the retention rates that is apparent when comparing Figure 33 to Figure 34 is the relationship of the retention rates to the treatment-control group ratio: When the ratio was 1:1, the retention of clusters was typically at its largest while the retention of the treatment group was typically at its smallest. The opposite relationship was true for the 1:9 condition, where the percent clusters retained was smallest and percent treatment group retained was largest. These relationships are related to the number of available matches for each treatment group member. When there are many control members with whom to match a treatment group member, the likelihood of finding an acceptable match among the controls is higher; therefore, the likelihood of retaining each treatment group member is also higher.

The cluster-level retention rate is also related to the number of treatment group members. In the 1:9 ratio condition, very few treatment group members are present in

each cluster; therefore, only limited opportunities are available to make treatment-control matches. When using stratification, however, it is far more likely that at least one control or treatment group member will be retained in each cluster than when matching methods are implemented.

Percent Significant γ_{10} across Propensity Score Adjustment Methods

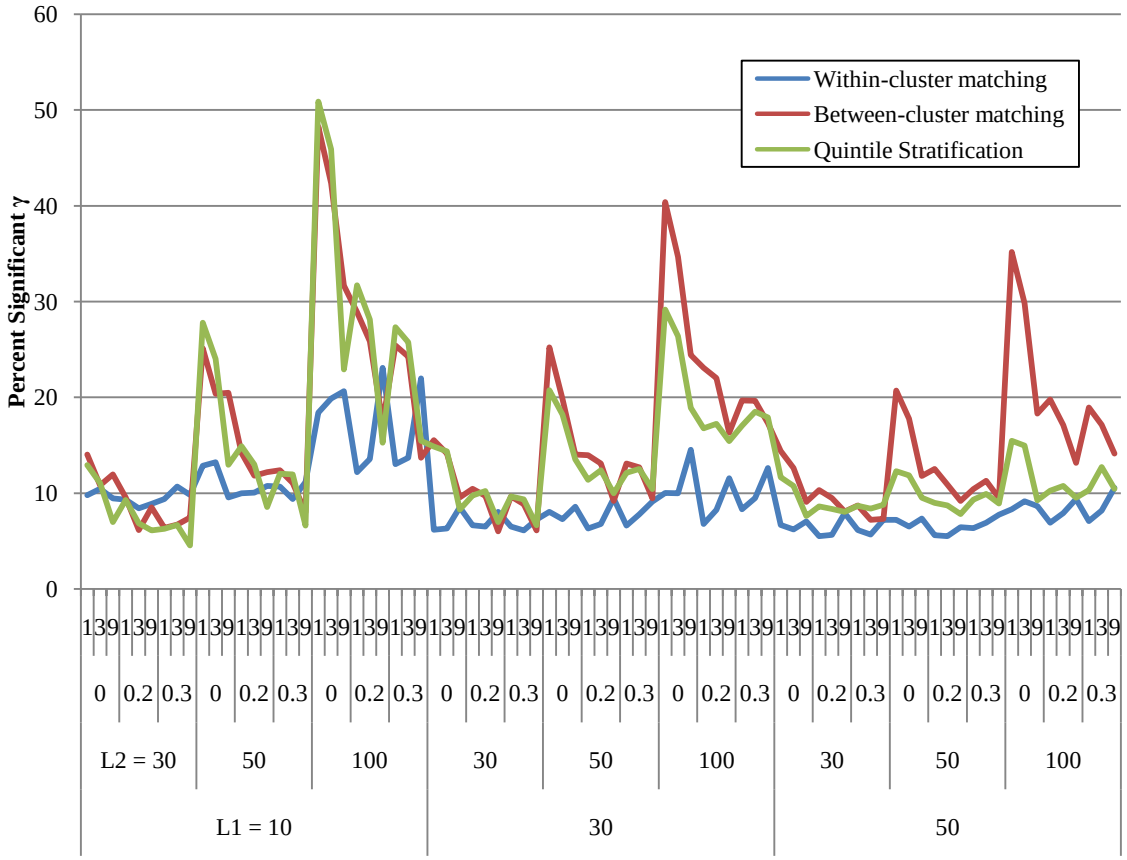
The average number of significant γ_{10} across Xs and across conditions was smallest for within-group matching (9.36), followed by stratification (13.92), and between-group matching (16.02). The average percent significant γ_{10} across Xs for within group matching ranged from 5.5 to 23.1, for between-group matching from 6.0 to 48.3, and for stratification from 4.6 to 50.9. Although within-group matching consistently retained smaller percentages of the sample than did the other adjustment methods, it also maintained the most consistent and smallest average percent significant γ_{10}. This finding could indicate successful balance or be a result of the low power to detect differences due to the small effective sample size. Stratification retained the largest sample but shows the greatest variability in percent significant γ_{10} across conditions. Findings indicate that between-group matching was consistently the moderate performer in treatment group retention and in variability in percent significant γ_{10}; however, between-group matching showed the highest average percent significant γ_{10} compared to other propensity score adjustment methods.

When sample size was smallest (10 individuals nested within 30 clusters), all adjustment methods showed similar percentages of significant γ_{10}, which were below 15%. All methods showed increases in percent significant γ_{10} as the number of clusters increased to 50 and to 100. Within-cluster matching maintained fewer significant γ_{10} in

most conditions, only being surpassed in performance by other methods when cross-level interactions were greatest and the difference in the number of treatment to control group members was greatest. As noted in the previous paragraph, however, the apparent advantage of within-cluster matching as indicated by lower percent significant γ_{10} may result from the inability to detect differences due to low sample size.

When sample size at level-1 was equal to 30 and 50, within-cluster matching consistently showed the smallest percentage significant γ_{10} and the smallest variation, remaining below 10% except when level-2 sample size was greatest and the treatment-control ratio was most imbalanced. In conditions where the level-1 sample size was 30 and 50, stratification showed a smaller percent significant γ_{10} than did between-group matching across simulation conditions. Much of the variation in the percent significant γ_{10} was not apparent once level-1 sample size reached 50; while the percent significant γ_{10} showed wide variation across conditions for between-group matching. The mean values for the percent significant γ_{10} across X are provided in Figure 35 across conditions.

Figure 35: Percentage of Significant γ_{10} across Simulation Conditions with a Propensity Score Estimated using Model 1



Percent Significant τ_{11} across Propensity Score Adjustment Methods

The average percent of significant τ_{11} across covariates and conditions for within-group matching was 10.2, for between-group matching was 5.9, and for stratification was 17.5. The average percent significant τ_{11} across Xs for within-group matching ranged from 0.7 to 26.5 (range = 25.8), for between-group matching from 0.8 to 20.0 (range = 19.2) and for stratification from 0.7 to 84.2 (range = 83.5).

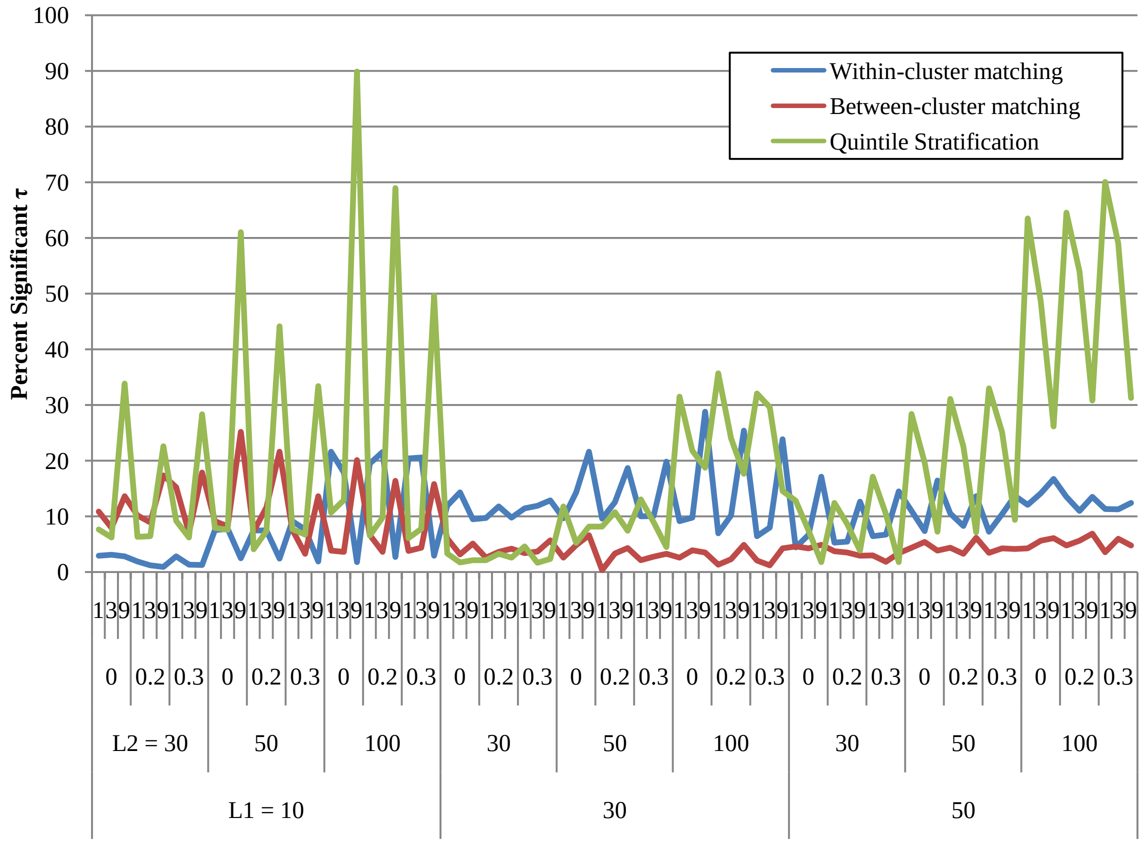
Results of stratification into quintiles indicated the greatest variability across conditions in the percent of significant τ_{11}. When the pool of control individuals was

large in comparison to the treatment group individuals, stratification tended to perform as well as other propensity score adjustment methods. Stratification adjustment maintained the percent of significant τ_{11} to less than 10%, other than when sample size at level-2 was 100. When the treatment-control ratio was 1:1, however, stratification methods resulted in the greatest percent of significant τ_{11} compared to other methods.

The application of propensity scores estimated from a multilevel model to between-cluster matching maintained the smallest percent of significant τ_{11} compared to other methods when the level-1 sample size was 30 and 50. When the level-1 sample size was smallest (n=10), however, the percent of significant τ_{11} fluctuated widely with the treatment-control group ratio, although not as widely as when stratification was used. Within-cluster matching maintained the smallest percent of significant τ_{11} relative to other propensity score adjustment methods in the smallest sample-size conditions, when level-1 sample size was 10 and level-2 sample size was 30 and 50. The percent of significant τ_{11} resultant from within-cluster matching was moderate in other sample conditions.

All methods maintained relatively low percent of significant τ_{11} when sample size at level-1 and at level-2 was equal to 30. The mean values for the percent significant τ_{11} across Xs are provided in Figure 36 across conditions.

Figure 36: Percentage of Significant τ_{11} Across Simulation Conditions with a Propensity Score Estimated using Model 1

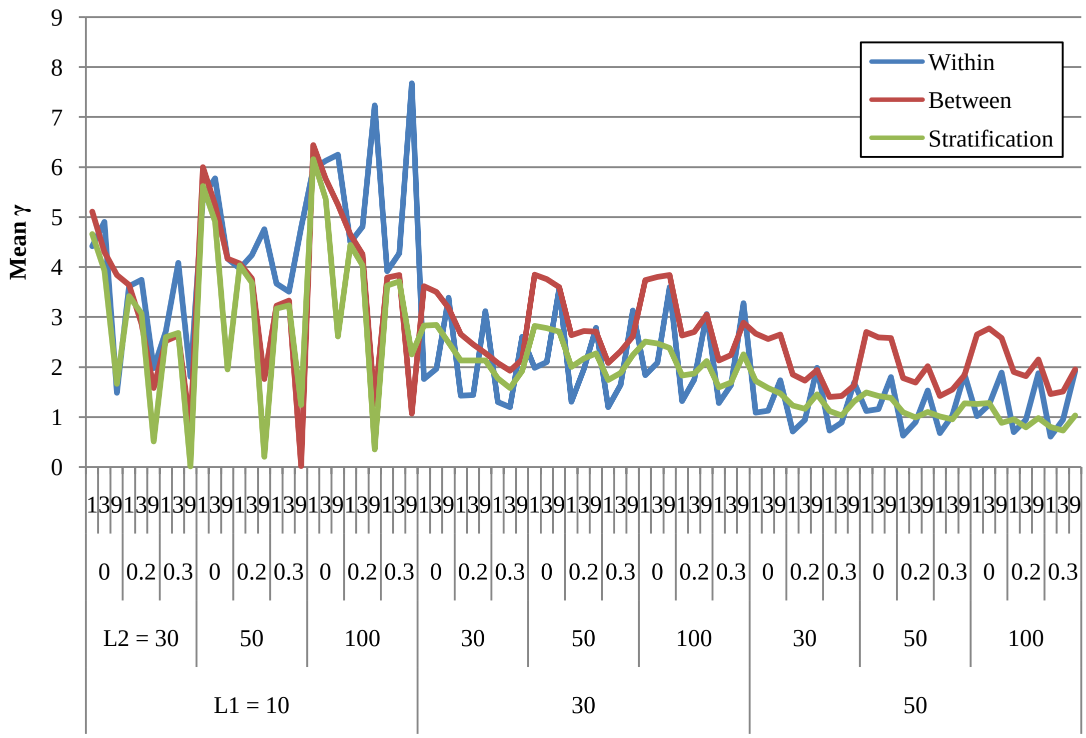


Mean γ_{10} and Mean τ_{11} across Propensity Score Adjustment Methods

The mean values of γ_{10} across the three adjustment methods show similar patterns across simulation conditions. The patterns from between-cluster matching and quintile stratification are nearly parallel once sample size at level-1 reaches 30. All methods show great variability when sample size at level-1 was 10. Considering the fact that the estimation model for this propensity score is not misspecified (e.g., it included all relevant predictors), the variance in the balance can likely be attributed to small sample size. The greater imbalance in the 1:9 condition when using within-cluster matching and smaller imbalance when using between-cluster matching and quintiles stratification is

apparent in Figure 37 below, especially when sample sizes were smallest. When forced to match within-clusters, the 1:9 condition provided no benefit to finding close matches, actually resulting in larger mean values γ_{10} of compared to 1:1 and 1:3 ratio conditions. When matching was not limited to within-clusters, however, the greater number of control group members versus treatment group members show clear benefit when sample size is small.

Figure 37: Mean Values of γ_{10} for X_1 across Simulation Conditions

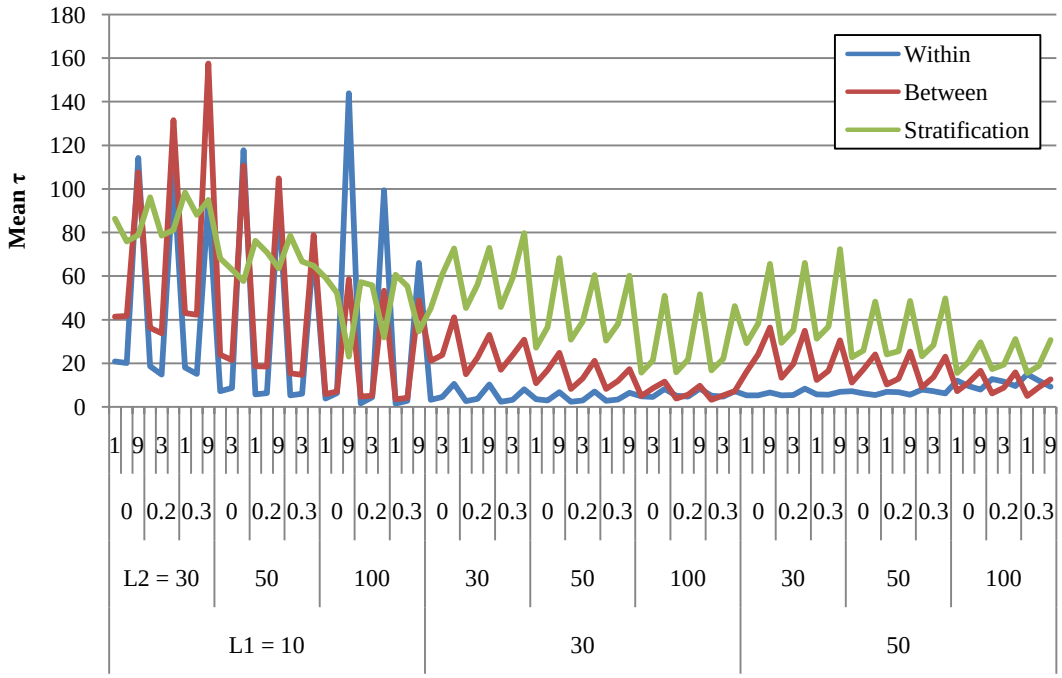


The variance in the balance across clusters shows the greatest fluctuation when sample size at level-1 is smallest. This fluctuation is primarily limited to the two matching adjustment methods, which are greatly influenced by the treatment-control ratio conditions. When more treatment group members are present per cluster, the likelihood that both treatment group and control group members will be present in each cluster increases. Subsequently, the variability in balance across cluster is smallest when the

ratio is 1:1 and 1:3 versus 1:9. Once sample size at level-1 reaches 30, however, much of this fluctuation disappears in the matching methods.

Once sample size at level-1 does reach 30, within-cluster matching results in very small variability in the balance achieved per cluster. Between-cluster matching performs as well as within-cluster matching when the number of clusters is large (e.g., 100). Quintile stratification shows the greatest variability in balance compared to other adjustment methods once sample size at level-1 reaches 30.

Figure 38: Mean Values of τ_{11} for X_1 across Simulation Conditions and Adjustment Method with a Propensity Score Estimated using Model 1



Standardized Mean Difference across Propensity Score Adjustment Methods

Across propensity score adjustment methods, an increase in level-1 sample size was related to a decrease in the SMD. This relationship was strongest in the within-cluster matching condition and weakest in the quintile stratification condition. The

relationship of the treatment-control ratio was also most pronounced in the within-cluster matching condition, in which greater numbers of control members relative to treatment control members resulted in greater SMD. In the other propensity score adjustment methods, between-cluster matching and stratification, the relationship of treatment-control ratio and SMD was trivial. Table 28 below presents the average SMD across covariates, for each adjustment method per level-1 sample size and treatment-control ratio.

Table 28:

Average SMD across Covariates per Level-1 Sample Size and Treatment-Control Ratio per Propensity Score Adjustment Method for a Propensity Score Estimated using Model 1

Level 1 Sample Size	Treatment-Control Ratio	Propensity Score Method		
		Within-Cluster Matching	Between-Cluster Matching	Quintile Stratification
10	1:1	52.3	32.6	45.2
	1:3	51.9	30.8	41.3
	1:9	70.0	31.3	44.6
30	1:1	20.3	22.6	35.1
	1:3	22.9	24.2	35.5
	1:9	35.1	27.1	34.6
50	1:1	14.9	18.4	30.6
	1:3	16.6	20.0	31.3
	1:9	23.9	23.5	32.6

Across propensity score adjustment methods, an increase in level-2 sample size also was related to a decrease in the SMD. As with level-1 sample size, this relationship was most pronounced when applying the propensity scores to within-cluster matching. Table 29, below, presents values for the average SMD across covariates for each propensity score adjustment method per Level-2 sample size and treatment-control ratio.

Table 29:

Average SMD across Covariates per Level-2 Sample Size and Treatment-Control Ratio per Propensity Score Adjustment Method for a Propensity Score Estimated using Model 1

		Propensity Score Method		
Level 2 Sample Size	Treatment:Control Ratio	Within-Cluster Matching	Between-Cluster Matching	Quintile Stratification
10	1:1	37.31	27.08	42.61
	1:3	38.33	27.68	41.34
	1:9	55.09	31.54	42.60
30	1:1	28.83	24.19	36.69
	1:3	30.17	24.78	35.82
	1:9	41.51	26.92	37.08
50	1:1	21.31	22.31	31.55
	1:3	22.81	22.54	30.88
	1:9	32.40	23.43	32.13

Performance of Propensity Scores Estimated using Logistic Regression

The results related to balance accomplished through application of the two propensity score estimation models that used logistic regression were similar. The two propensity score estimation models were: Model 2, which included covariates X_1, X_2, X_3, and W, and Model 3, which included X_1, X_2, and X_3 but not W. These two models are discussed together, with emphasis on any clear divergence in their results.

Within-cluster matching

Balance Achievement in Predictor Covariates: Percent Significant γ10

Overall, the percent significant γ_{10} was small when applying within-cluster matching with propensity scores estimated from logistic regression. This finding indicates that the relationship between the treatment group assignment and each covariate was successfully removed across the sample as a whole. The percent of the treatment group retained after matching, however, was also small, with a mean of 37% across cells.

Sample Size at Level-1 and Level-2. A negative relationship exists between the percent significant γ_{10} and the level-1 sample size; however, the change in this percent is small within each level-1 sample-size condition. The relationship of percent significant γ_{10} and level-2 sample size is trivial. The results for each of these propensity scores are nearly identical per condition, as illustrated in Table 30

Table 28 below.

Table 30:

Percentage Significant γ_{10} Resultant from Within-Cluster Matching using Propensity Scores Estimated from Model 2 and Model 3 per Sample Size Condition

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
				Model 2						
30	2.7	0.2	0.0	3.3	0.2	0.0	2.3	0.1	0.0	1.0
50	1.7	0.1	0.0	2.2	0.1	0.0	1.9	0.2	0.0	0.7
100	1.9	0.1	0.0	2.2	0.1	0.0	1.9	0.3	0.5	0.8
Mean	2.1	0.1	0.0	2.6	0.2	0.0	2.0	0.2	0.2	0.8
				Model 3						
30	2.6	0.1	0.0	2.5	0.1	0.0	2.2	0.1	0.0	0.9
50	2.0	0.1	0.0	2.4	0.1	0.0	2.2	0.2	0.1	0.8
100	1.8	0.1	0.0	1.9	0.2	0.0	2.0	0.5	1.0	0.8
Mean	2.1	0.1	0.0	2.3	0.2	0.0	2.1	0.3	0.4	0.8

Cross-Level Interaction. The strength of the cross-level interaction in the generating sample was not related to the percent significant γ_{10}. Within each sample-size condition, the percent significant γ_{10} remained constant. With a sample size at level-1 of 30 and 50, the percent significant γ_{10} remained below 0.5% across conditions (see Table 31).

Table 31:

Percentage Significant γ_{10} for X_1 per Cross-Level Interaction Resultant from Within-Cluster Matching using Propensity Scores Estimated from Model 2 and Model 3

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	2.9	1.7	1.7	0.1	0.1	0.1	0.0	0.0	0.0	0.7
0.2	2.7	1.8	2.1	0.3	0.1	0.2	0.0	0.0	0.0	0.8
0.3	2.4	1.7	1.9	0.2	0.1	0.1	0.0	0.0	0.0	0.7
	Model 3									
0.0	2.9	1.7	1.5	0.0	0.1	0.1	0.0	0.0	0.0	0.7
0.2	2.5	1.9	1.7	0.1	0.0	0.1	0.0	0.0	0.0	0.7
0.3	2.5	2.4	2.3	0.1	0.1	0.1	0.1	0.0	0.0	0.8

Treatment-Control Group Ratio. The mean percent significant γ_{10} shows a non-linear relationship with the treatment-control group ratios. The mean percent significant γ_{10} is smallest when the ratio is 1:3, largest when the ratio is 1:9, and moderate when the ratio is 1:1. This pattern is consistent across cross-level interaction conditions (see Figure 41) and level-2 sample sizes (see Figure 40). This relationship is also evident when the level-1 sample size is smallest (n=10). The mean percent significant γ_{10}, however, decreases to approximately 0 when level-1 sample sizes are 30 and 50 (see Figure 39).

Figure 39: Mean Percent Significant γ_{10} for X_1 per treatment-control ratio and level-1 sample size

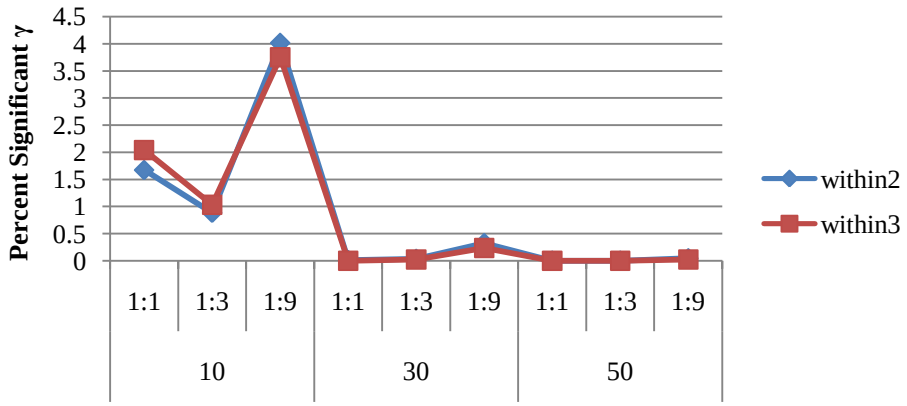


Figure 40: Mean Percent Significant γ_{10} for X_1 per Treatment-Control Ratio and Level-2 Sample Size

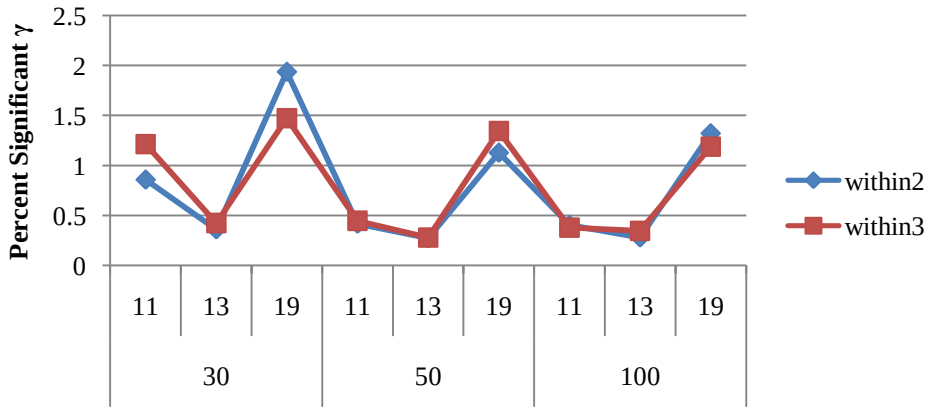


Figure 41: Mean Percent Significant γ_{10} for X_1 per Treatment-Control Ratio and Cross-Level Interaction

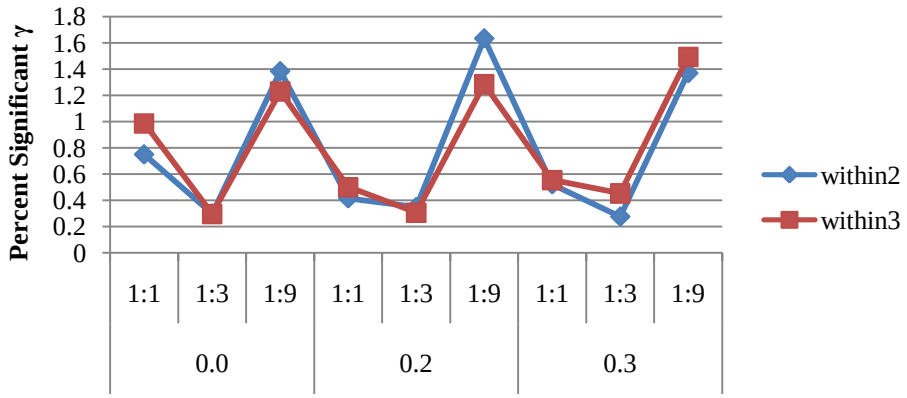
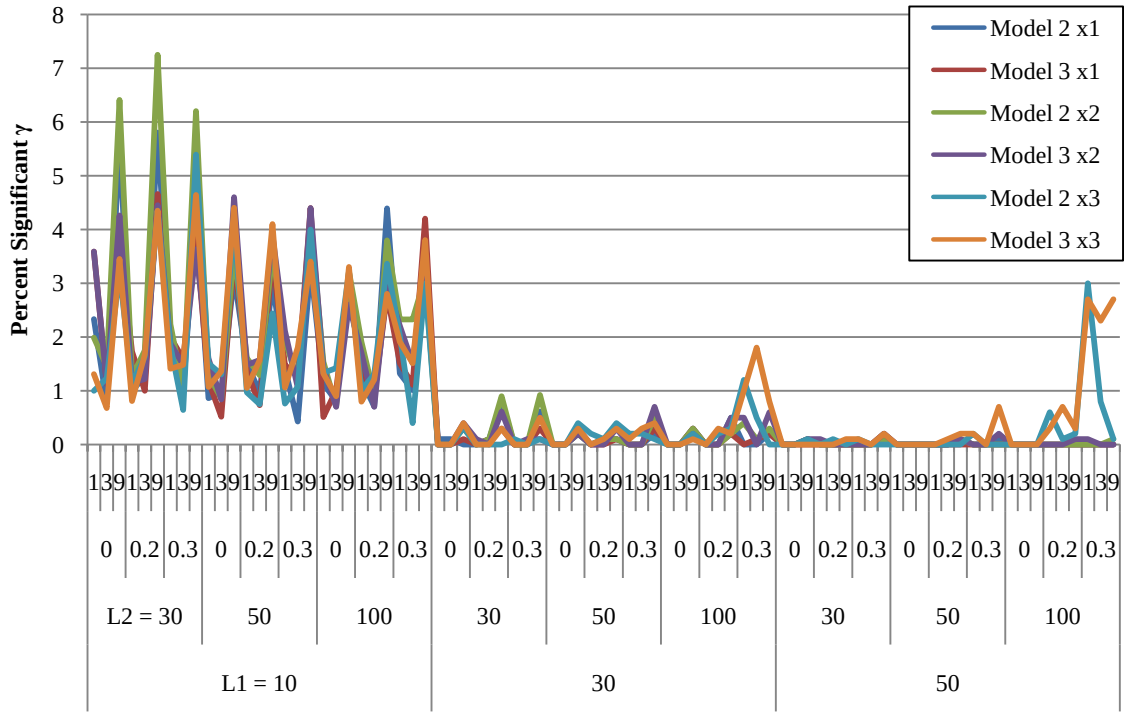


Figure 42: Percent of Significant γ_{10} for X_1, X_2, and X_3 across Simulation Conditions for Within-Cluster Matching using Propensity Scores Estimated using Model 2 and Model 3



Variation in Balance Achievement within Clusters: Percentage Significant τ11

Sample Size at Level-1 and Level-2. The percentage of significant τ_{11} is positively related to sample size at both level-1 and level-2: as sample size increases, the percentage of significant τ_{11} increases. The range in percentage of significant τ_{11} is broad, equal to 12% at the smallest sample-size conditions and 100% at the largest. The average percent of significant τ_{11} per sample-size condition are presented in Table 32.

Table 32:

Average Percentage of Significant τ_{11} Resultant from Within-Cluster Matching using Propensity Scores Estimated using Model 2 and Model 3 per Sample Size

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
					Model 2					
30	12.0	46.8	72.7	22.7	80.7	95.8	32.7	92.7	99.7	61.8
50	14.8	57.9	81.9	29.8	90.6	99.3	48.4	98.8	100.0	69.1
100	16.1	68.8	90.0	45.2	98.0	100.0	65.8	100.0	100.0	76.0
Mean	14.3	57.8	81.5	32.6	89.8	98.4	48.9	97.2	99.9	68.9
					Model 3					
30	12.5	46.9	72.9	21.5	81.6	96.4	32.5	93.4	99.6	61.9
50	15.0	58.1	81.5	30.3	91.9	99.4	49.0	98.9	100.0	69.4
100	16.0	68.4	88.8	45.4	98.5	100.0	66.0	100.0	100.0	75.9
Mean	14.5	57.8	81.1	32.4	90.6	98.6	49.2	97.4	99.9	69.1

Cross-Level Interaction. Findings indicate a negative relationship exists between the strength of the cross-level interaction and the percentage of significant τ_{11}: As the strength of the cross-level interaction increases, the percentage of significant τ_{11} decreases. The values for percentage of significant τ_{11} resulting from within-cluster matching using Model 2 are nearly identical to those of Model 3 (see Table 33).

Table 33:

Average Percentage of Significant τ_{11} for X_1 per Cross-Level Interaction Resultant from Within-Cluster Matching using Propensity Scores Estimated using Model 2 and Model 3

$\rho_{(WX)Z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	13.6	15.6	19.8	54.7	68.1	78.6	79.3	89.1	95.7	57.2
0.2	11.6	16.0	15.1	47.0	57.9	70.6	71.6	82.3	90.8	51.4
0.3	10.9	12.8	13.4	38.8	47.6	57.3	67.1	74.3	83.4	45.1
	Model 3									
0.0	12.8	16.4	18.8	54.4	67.7	79.4	79.6	88.7	96.1	57.1
0.2	12.7	15.0	14.7	47.0	58.5	70.2	72.1	81.7	89.2	51.2
0.3	12.0	13.6	14.4	39.2	48.2	55.6	67.0	74.2	81.2	45.0

Treatment-Control Group Ratio. When matching within clusters, an approximately linear relationship is apparent between the proportion of treatment group members to control group members and the percentage of significant τ_{11}. Smaller

percentages of significant τ_{11} are apparent when the treatment-control group ratio is 1:9 versus 1:3 versus 1:1. This pattern is consistent across sample-sizes and strength of the cross-level interaction, indicating an absence of any interaction effects (see Figure 43, Figure 44, and Figure 45).

Figure 43: Mean Percentage of Significant τ_{11} per Treatment-Control Ratio and Level-1 Sample Size.

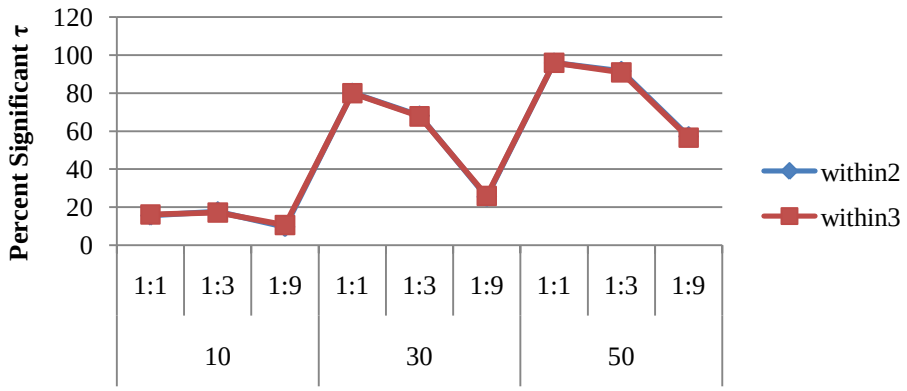


Figure 44: Mean Percentage of Significant τ_{11} per Treatment-Control Ratio and Level-2 Sample Size.

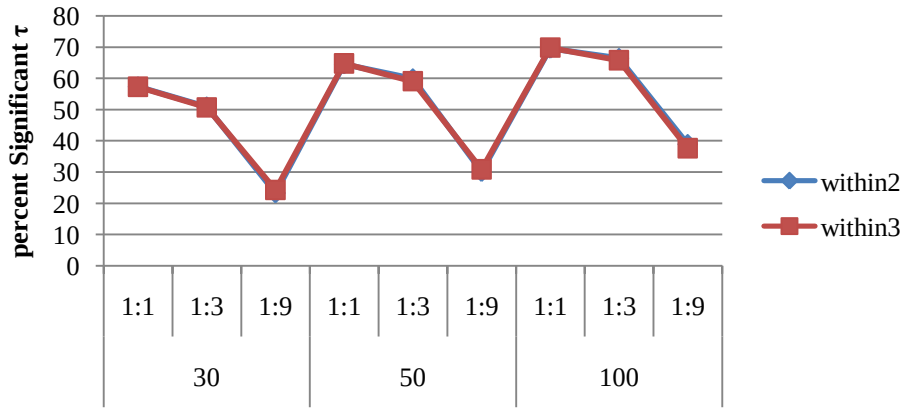
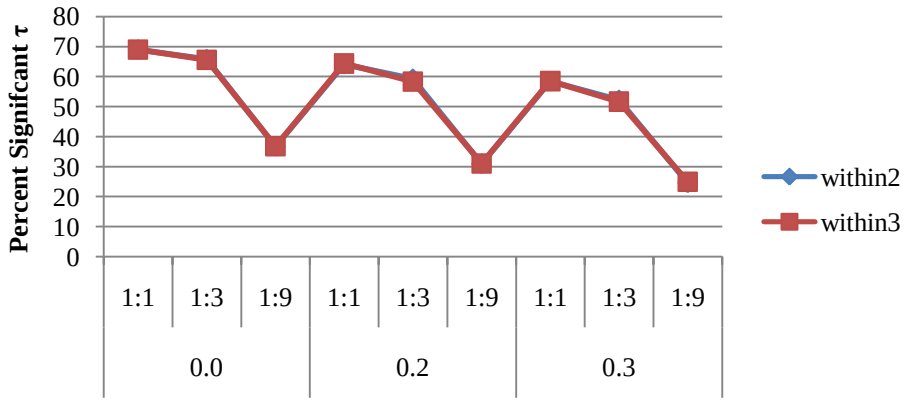


Figure 45: Mean Percentage of Significant τ_{11} per Cross-Level Interaction and Treatment-Control Group Ratio.

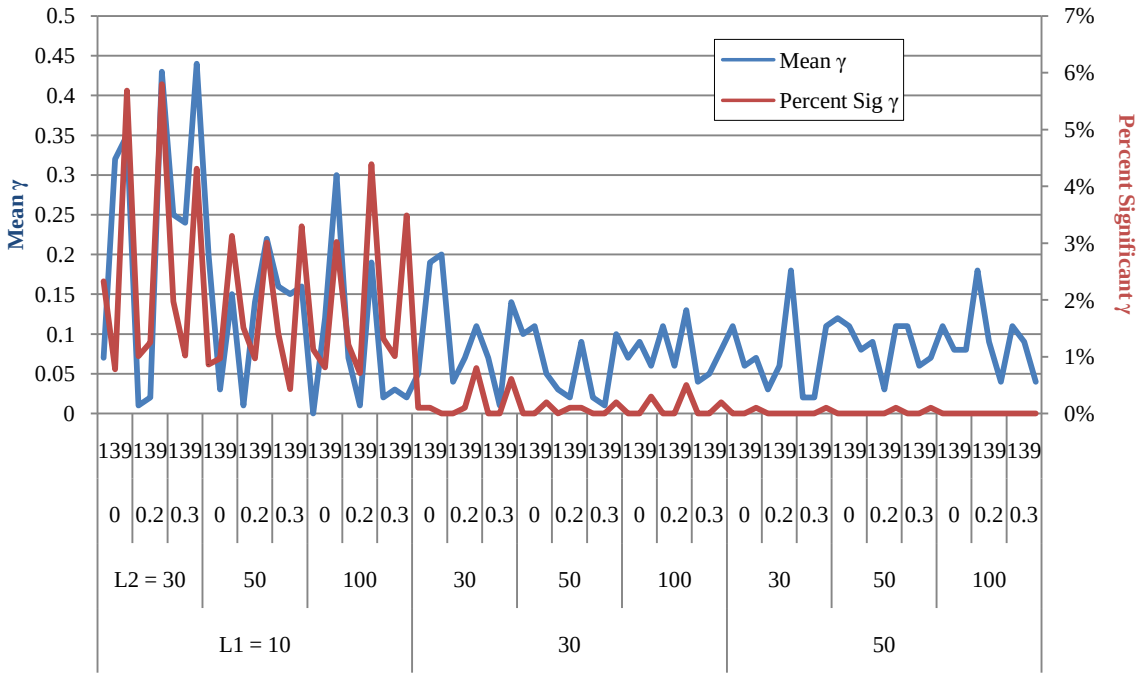


Percent Significant versus Mean Value of γ_{10} and τ_{11}

When applying a propensity score estimated through Model 2 to within-cluster matching, larger sample sizes result in lower percent significant γ_{10} and smaller mean values of γ_{10}. This finding supports the conclusion that the performance of this adjustment method to balance covariates across conditions improves with larger sample sizes.

The pattern resulting from a propensity score estimated from Model 2 (illustrated in Figure 46 below) was nearly identical to that estimated from Model 3, which did not include a cluster-level predictor. The similarities of the results from these two models are likely resultant from the method of propensity score application: within-cluster matching. Matching between individuals who are in the same cluster effectively controls for the cluster-level effects.

Figure 46: Overlay of Mean γ_{10} and Percent Significant γ_{10} for Within-Cluster Matching with Propensity Scores Estimated using Model 2



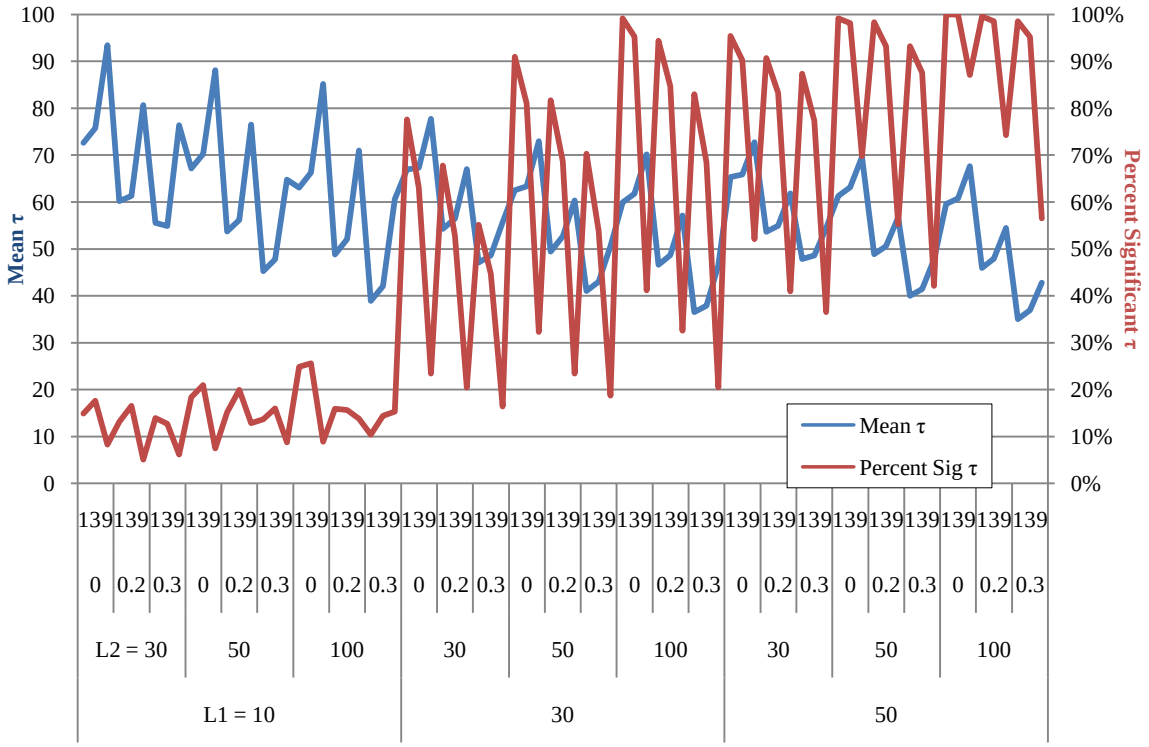
When matching is limited to individuals within each cluster, the mean value of τ_{11}, representing the variability in the balance across clusters, is smallest when sample size is largest. When considering the percent significant τ_{11}, however, statistically greater variability is apparent with larger sample sizes. These contradictory findings suggest that the increase in percent significant τ_{11} is related to increased power to detect the variance in the slopes associated with larger sample sizes rather than greater variances. This conclusion is supported by the fact that the treatment-control ratio condition that shows the smallest values for the percent significant τ_{11} (1:9 condition) also has the smallest number of treatment group members and subsequent overall sample size.

By contrast, the 1:9 condition also shows the largest mean values for τ_{11}. This finding suggests that, although more controls are available with which to match treatment

group members within each cluster, the matches do not result in better balance within the clusters. A potential explanation for this finding is that having more treatment group individuals (1:1 condition) likely provides better propensity score estimations and better matches than when there are fewer treatment group members (1:9 condition). This pattern in the findings is nearly identical for propensity scores estimated with a cluster-level predictor (Model 1) and those estimated without such a predictor (Models 2 and 3). In other words, when matching within-clusters, the presence of the cluster-level predictor has little bearing on the covariate balance achieved.

Another pattern that is of interest in the data is the smaller variance that is related to larger cross-level interactions. A possible explanation for this effect is related to the within-cluster matching mechanism. In this case, the effect of the cluster-level predictor (e.g., the cross-level interactions) had little bearing upon the quality of the matches. When estimating the propensity scores when $\rho_{(WX)Z} = 0$, including X_3 in the equation merely added noise to the estimates; however, when $\rho_{(WX)Z} = .2$ and .3, X_3 contributed to the propensity score which resulted in better matches and less variance in the balance across clusters.

Figure 47: Overlay of Mean τ_{11} and Percent Significant τ_{11} for Within-Cluster Matching with Propensity Scores Estimated using Model 2

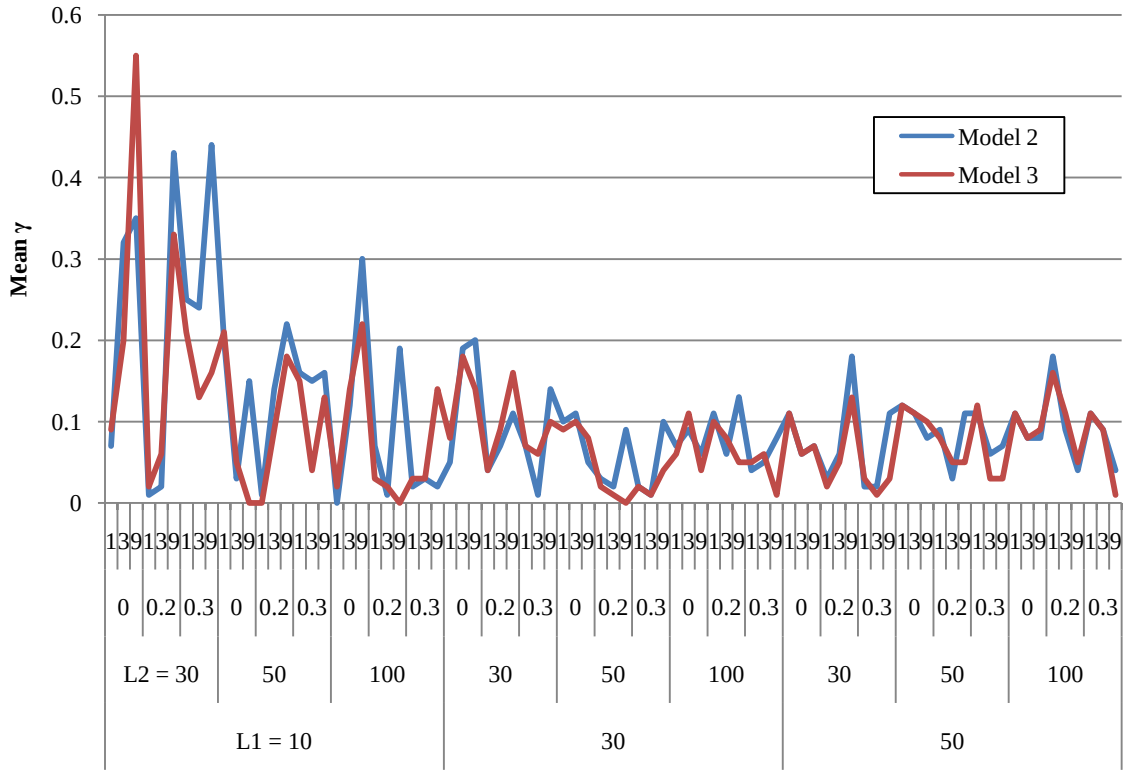


Comparison of Mean γ_{10} for Propensity Scores Estimated using Model 2 and Model 3

Although the values for the mean γ_{10} are similar for within-cluster matching using propensity scores estimated using Model 2 and Model 3, a few conditions show differences. The values are most dissimilar when sample size at level-1 is small; however, a clear pattern of the differences is not apparent. When sample size is smallest (L1 = 10, L2 = 30, Ratio 1:9), Model 2 shows larger mean values of γ_{10} with stronger cross-level interactions, whereas Model 3 shows the opposite relationship. This finding indicates that including a cluster-level predictor in the model results in increased imbalance when the sample size is too small to estimate adequately the influence of that predictor. Overall, both models appear to result in small mean values of γ_{10} in all

conditions, with values remaining below .6. A comparison of the mean values for τ_{11} for these two propensity score models is not presented because no differences were apparent.

Figure 48: Mean γ_{10} for Propensity Scores Estimated using Model 2 and Model 3 when applied to Within-Cluster Matching



Balance Achievement in Predictor Covariates: The Standardized Mean Difference

Sample Size at Level-1 and Level-2. The findings suggested by the mean SMD per sample-size condition supports the previous findings related to overall balance: As sample-size increases, the mean SMD on each covariate decreases. This pattern is nearly identical for propensity scores estimated with cluster-level covariates (Model 2) and without cluster-level covariates (Model 3).

Table 34:

Mean SMD Resultant from Within-Cluster Matching using Propensity Score Estimated using Model 2 and Model 3

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
					Model 2					
30	22.1	7.8	5.3	40.5	14.0	9.6	35.6	12.6	8.5	17.3
50	16.2	5.9	4.0	29.9	11.0	7.6	26.3	9.6	6.5	13.0
100	11.0	4.2	2.8	20.5	7.9	5.5	18.0	6.9	4.6	9.0
Mean	16.5	6.0	4.0	30.3	11.0	7.6	26.6	9.7	6.6	13.1
					Model 3					
30	21.5	7.7	5.2	39.4	14.3	9.8	34.5	12.5	8.5	17.0
50	16.0	5.8	4.0	29.3	11.3	7.9	26.1	9.7	6.6	13.0
100	10.9	4.1	2.8	20.2	8.2	6.1	17.9	6.9	4.8	9.1
Mean	16.1	5.9	4.0	29.6	11.3	7.9	26.1	9.7	6.6	13.0

Cross-Level Interaction. The relationship between the strength of the cross-level interaction and the mean SMD is small or nonexistent. This relationship is similar to that found between the cross-level interaction and percent significant γ_{10}.

Table 35:

SMD for X_1 per Cross-Level Interaction Resultant from Within-Cluster Matching using Propensity Scores estimated using Model 2 and Model 3

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	22.5	16.7	11.2	7.7	5.9	4.3	5.4	4.1	2.8	9.0
0.2	22.0	16.0	11.0	7.9	5.9	4.2	5.2	4.0	2.8	8.8
0.3	21.8	16.0	10.9	7.7	5.9	4.1	5.3	3.8	2.7	8.7
	Model 3									
0.0	21.7	16.5	11.1	7.8	5.9	4.2	5.4	4.1	2.9	8.8
0.2	21.1	15.7	11.0	7.7	5.8	4.1	5.1	4.0	2.8	8.6
0.3	21.6	15.7	10.7	7.5	5.8	4.0	5.0	3.8	2.7	8.5

Treatment-Control Group Ratio. A slight increase in the mean SMD is evident as the treatment-control group ratio changes from 1:1 to 1:3. In most cases, this increase in SMD is greater when the ratio changes from 1:3 to 1:9. Across conditions, the 1:9 ratio consistently results in the largest SMD between covariates. No interaction effect is apparent between this relationship and level-2 sample size or strength of the cross-level interaction. The relationship between treatment-control group ratio and SMD becomes less evident as the sample-size at level-1 increases (see Figure 49, Figure 50, and Figure 51).

Figure 49: Mean SMD per Level-1 Sample Size

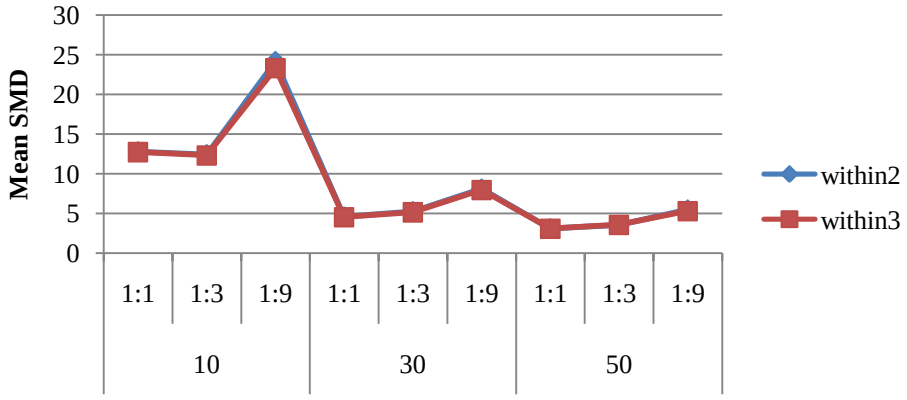


Figure 50: Mean SMD per Level-2 Sample Size

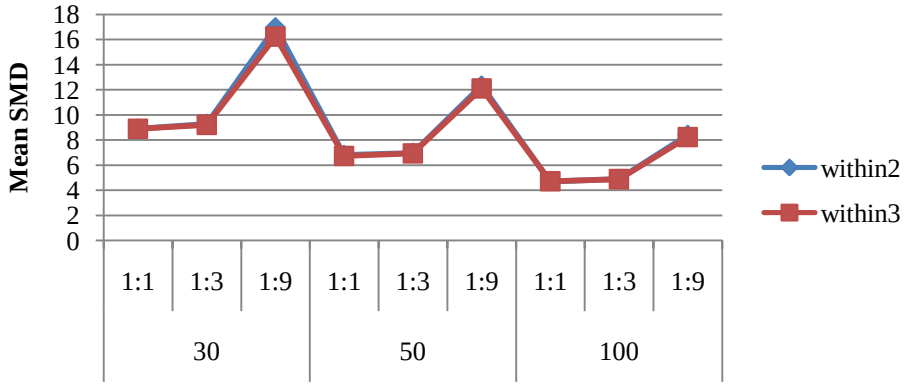


Figure 51: Mean SMD per Cross-Level Interaction

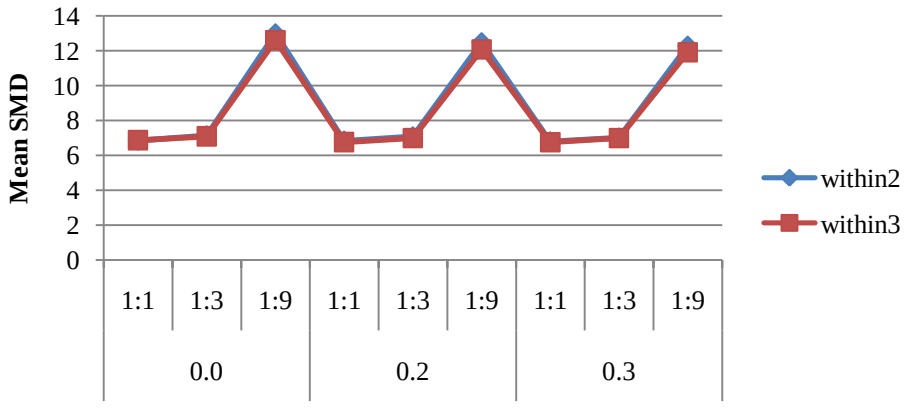
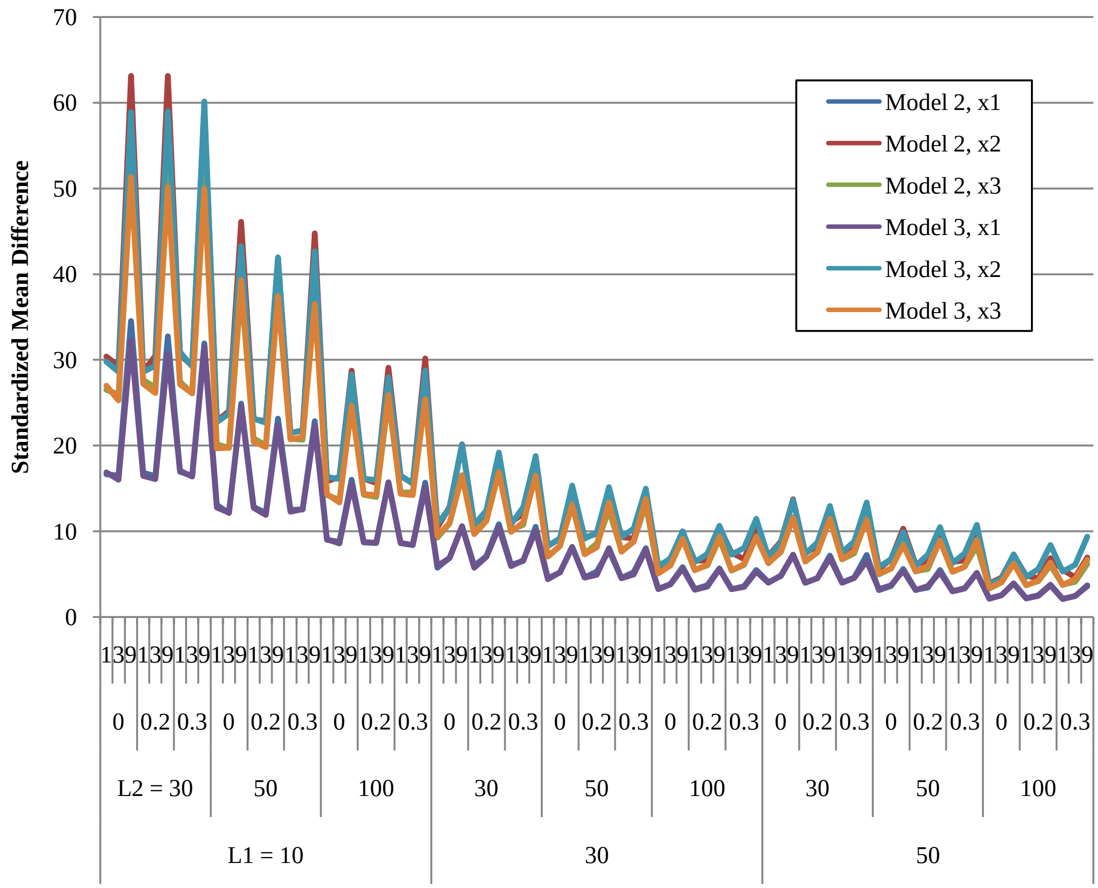


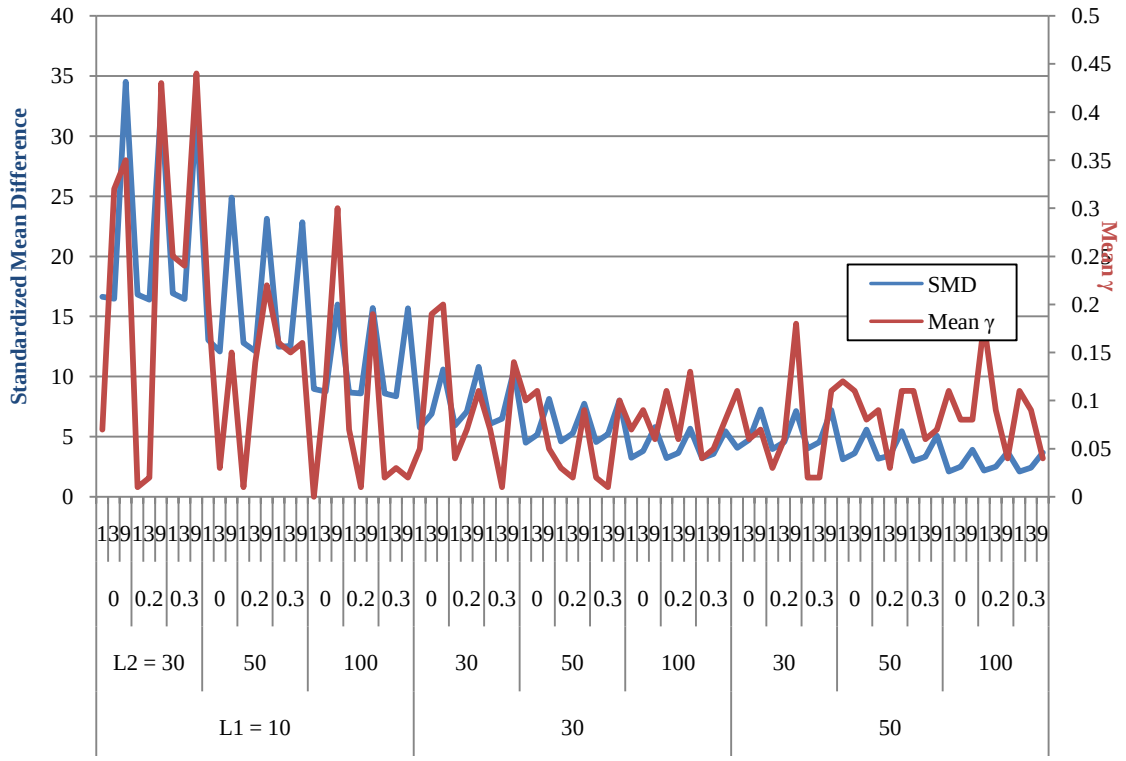
Figure 52: Mean SMD for X_1, X_2, and X_3 across Simulation Conditions for Within-Cluster Matching using a Propensity Score estimated using Model 2 and Model 3



SMD vesus Mean γ_{10}

Although the SMD shows less fluctuation across conditions than does the mean γ_{10}, their overall patterns are similar: overall balance is better with larger sample sizes and with treatment-control ratios of 1:9 (vs. 1:3 and 1:1). This pattern is similar for propensity scores estimated with and without a cluster-level predictor. Values for SMD and Mean γ_{10} resulting from application of a propensity score estimated using Model 2 are presented in Figure 53 below.

Figure 53: Overlay of SMD and Mean γ_{10} for Within-Cluster Matching with Propensity Scores Estimated using Model 2



Between-cluster matching

Balance Achievement in Predictor Covariates: Percentage Significant γ_{10}

Overall, between-cluster matching using propensity scores that are estimated using a logistic model, with or without the inclusion of a cluster-level predictor, resulted in very few significant γ_{10} across sample-size conditions.

Sample Size at Level-1 and Level-2. Under the majority of sample-size conditions, the percentage of significant γ_{10} was approximately 0. For the variable X_1, the results indicate that a positive relationship exists between percentage of significant γ_{10} and both level-1 sample size and level-2 sample size: As the sample size increases, the percentage of significant γ_{10} increases. The percentage of significant γ_{10} is slightly higher for results

from propensity score estimation Model 3 versus Model 2. These values per sample-size condition are presented in Table 36.

Table 36:

Average Percentage of Significant γ_{10} for X_1 Resultant from Between-Cluster Matching using a Propensity Scores Estimated using Model 2 and Model 3

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
					Model 2					
30	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
50	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
100	0.0	0.6	3.1	0.0	0.0	0.1	0.0	0.0	0.0	0.4
Mean	0.0	0.2	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
					Model 3					
30	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
50	0.0	0.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.1
100	0.0	1.5	9.4	0.0	0.1	0.3	0.0	0.0	0.0	1.3
Mean	0.0	0.5	3.2	0.0	0.0	0.1	0.0	0.0	0.0	0.4

Cross-Level Interaction. When matching between-clusters using a logistic regression model for propensity score estimation, the strength of the cross-level interaction has a negative relationship with the percentage of significant γ_{10}: As the strength of the cross-level interaction increases, the percentage of significant γ_{10} decreases. This relationship is only apparent with larger sample sizes, especially when

level-2 sample size is largest (N=100). These values per sample-size condition are presented in Table 36.

Table 37:

Average Percentage of Significant γ_{10} for X_1 per Cross-Level Interaction Resultant from Between-Cluster Matching using Regression Models for Propensity Score Estimation per Level-1 and Level-2 Sample Sizes

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.2	6.1	0.8
0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	2.0	0.3
0.3	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	1.1	0.2
	Model 3									
0.0	0.0	0.0	0.0	0.0	0.1	2.3	0.1	0.4	12.5	1.7
0.2	0.0	0.0	0.0	0.0	0.0	1.7	0.0	0.4	9.8	1.3
0.3	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.1	5.9	0.7

Treatment-Control Group Ratio. The relationship of the treatment-control group ratio and percent significant γ_{10} shows little variability across sample size conditions at level-1, sample size conditions at level-2, and cross-level interactions. This similarity in patterns suggests that no interaction effects exist across other conditions and this relationship between treatment-control group ratio and percentage significant γ_{10}. This

relationship, however, is not the same for the propensity score estimated from Model 2 versus Model 3. For both propensity score estimation models, the 1:9 ratio condition showed smaller percent significant γ_{10}. With Model 2, the change in the percent significant γ_{10} was approximately linear: The 1:9 ratio condition showed fewer significant γ_{10} than the 1:3 ratio and 1:1 ratio. When propensity scores were estimated using Model 3, the moderate ratio condition (1:3) tended to show larger percent significant γ_{10} than did the 1:1 and 1:9 ratio conditions. These patterns are illustrated in Figure 54, Figure 55, and Figure 56.

Figure 54: Mean percent significant γ_{10} per treatment-control ratio and level-1 sample size.

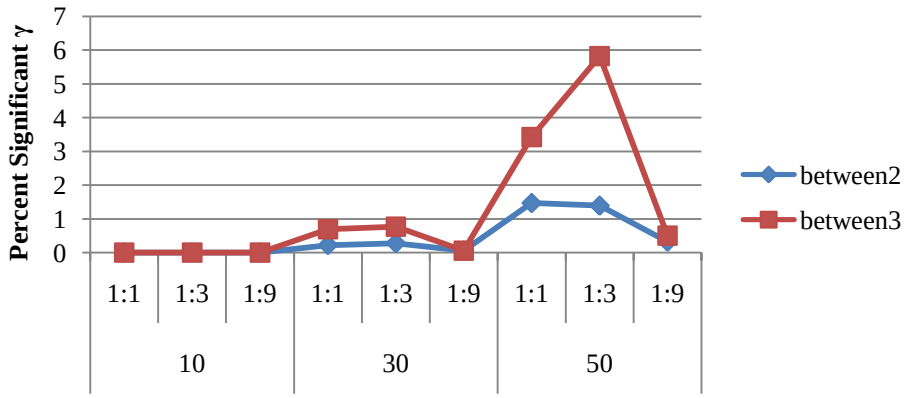


Figure 55: Mean percent significant γ_{10} per treatment-control ratio and level-2 sample size.

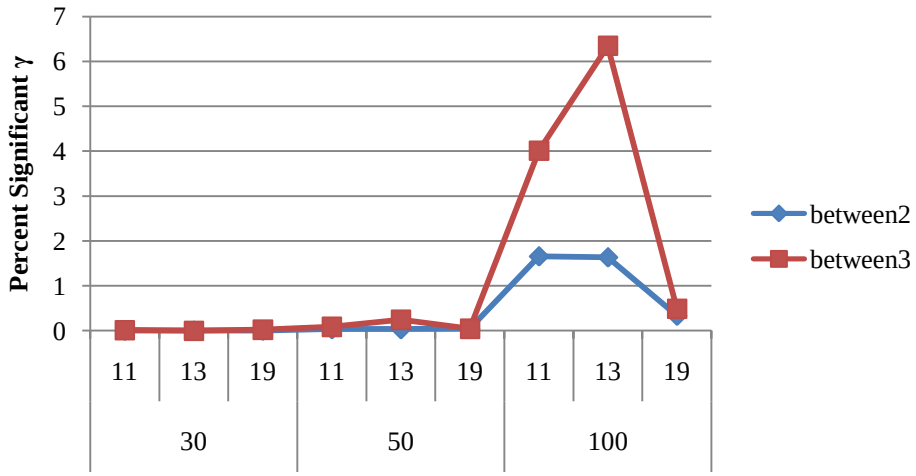


Figure 56: Mean percent significant γ_{10} per cross-level interaction and treatment-control group ratio.

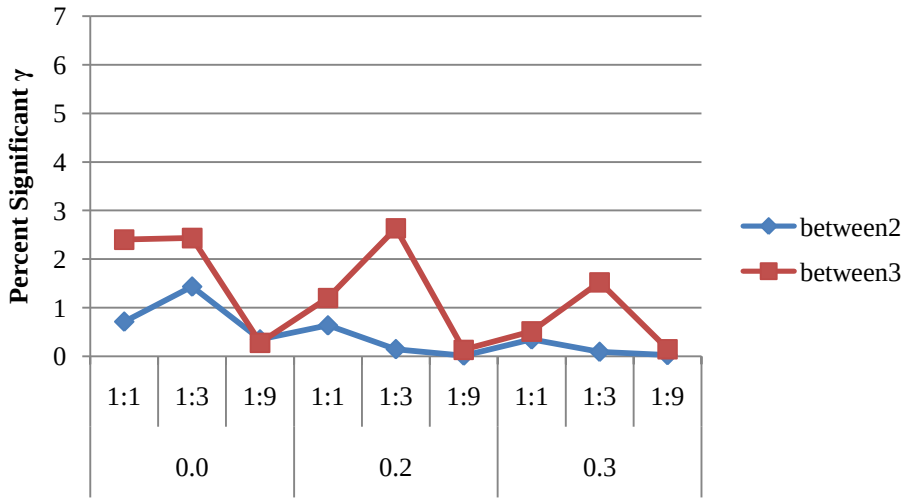
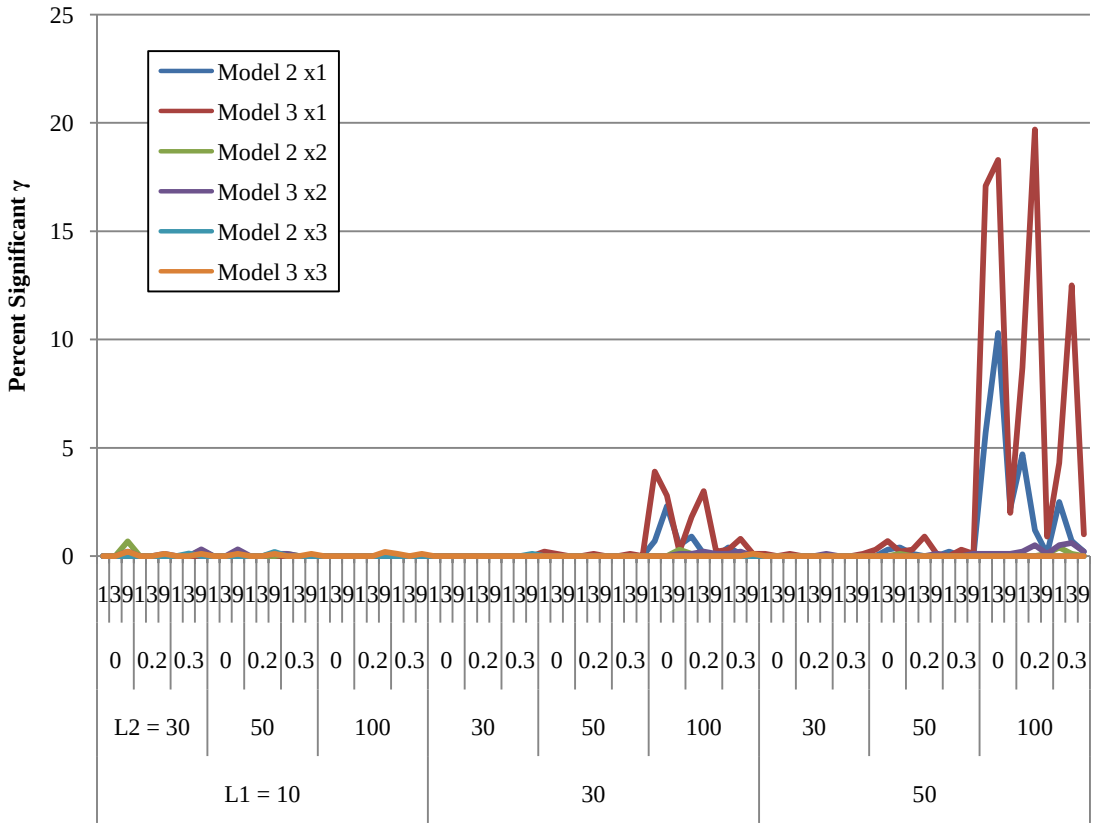


Figure 57: Percentage of Significant γ_{10} for X_1, X_2, and X_3 across Simulation Conditions for Between-Cluster Matching using Propensity Scores estimated using Model 2 and Model 3



Variation in Balance Achievement within Clusters: Percent Significant τ_{11}

Sample Size at Level-1 and Level-2. Results from the examination of the percent significant τ_{11} suggest that the variance in the balance across clusters was consistently very large when between-cluster matching was used to balance covariates with propensity scores estimated using logistic models. The percent significant τ_{11} showed no meaningful differences between those attained using a propensity score estimated using Model 2 versus Model 3. The relationship between sample size and the percent significant τ_{11} was positive: As the sample-size increased at level-1 or level-2, the percent

significant τ_{11} increased. The mean percent significant τ_{11} per sample-size condition is presented in Table 38.

Table 38:

Average Percentage of Significant τ_{11} Resultant from Between-Cluster Matching using a Multilevel Propensity Score Estimation Model per Level-1 and Level-2 Sample Size

Level-2 Sample Size	X_1 Level-1 Sample Size			X_2 Level-1 Sample Size			X_3 Level-1 Sample Size			Mean
	10	30	50	10	30	50	10	30	50	
Model 2										
30	62.3	94.8	99.5	69.5	96.7	99.8	74.1	97.8	99.9	88.3
50	75.6	99.3	100.0	80.3	99.5	100.0	83.5	99.9	100.0	93.1
100	83.9	100.0	100.0	88.0	100.0	100.0	91.3	100.0	100.0	95.9
Mean	73.9	98.0	99.8	79.3	98.8	99.9	83.0	99.2	100.0	92.4
Model 3										
30	61.9	94.6	99.3	67.3	95.9	99.7	71.9	97.3	99.8	87.5
50	76.9	99.1	100.0	80.4	99.7	100.0	82.5	99.8	100.0	93.2
100	87.2	100.0	100.0	88.1	100.0	100.0	90.1	100.0	100.0	96.2
Mean	75.3	97.9	99.8	78.6	98.5	99.9	81.5	99.1	99.9	92.3

Cross-Level Interaction. The simulation results indicate that the strength of the cross-level interaction has no relationship with the percent significant τ_{11} when using between-cluster matching. The mean percentage significant τ_{11} for X_1 are presented in Table 39.

Table 39:

Average Percentage of Significant τ_{11} for X_1 per Cross-Level Interaction Resultant from Between-Cluster Matching using Propensity Scores Estimated using Model 2 and Model 3

$\rho_{(WX)Z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	64.3	76.8	88.0	94.9	99.0	100.0	99.3	100.0	100.0	91.4
0.2	61.6	75.7	82.4	94.5	99.4	100.0	99.7	99.9	100.0	90.4
0.3	61.0	74.2	81.2	95.2	99.4	100.0	99.6	100.0	100.0	90.1
Mean	62.3	75.6	83.9	94.8	99.3	100.0	99.5	100.0	100.0	90.6
	Model 3									
0.0	58.8	76.0	88.1	94.0	98.9	100.0	99.1	100.0	100.0	90.5
0.2	62.3	77.4	86.3	94.5	99.3	100.0	99.3	100.0	100.0	91.0
0.3	64.6	77.5	87.2	95.5	99.1	100.0	99.5	100.0	100.0	91.5

Treatment-Control Group Ratio. Findings from this study indicate that as the treatment-control group ratio becomes more imbalanced, the percent significant τ_{11} decreases. The ratio of 1:9 showed the smallest percent significant τ_{11} under most conditions. This pattern is not evident once sample size at level-1 reaches 30 where approximately all of the cells showed an average percent significant τ_{11} of 100%. No

interaction effects are evident between this relationship and sample-size conditions or the strength of the cross-level interactions (see Figure 58, Figure 59, and Figure 60).

Figure 58: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-1 sample size.

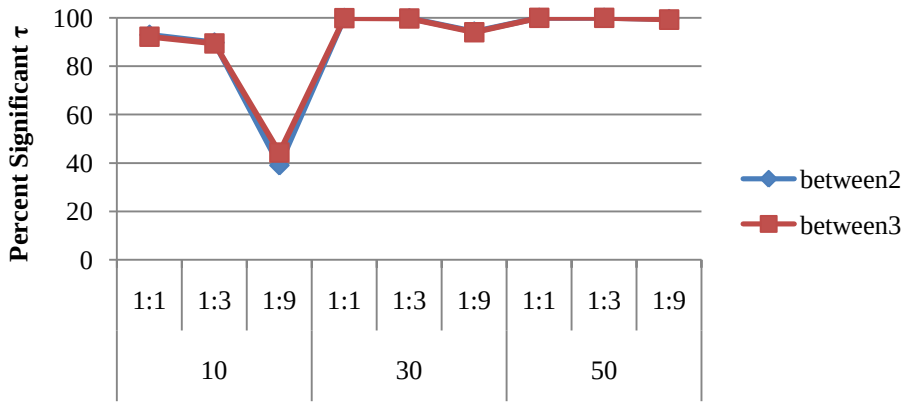


Figure 59: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-2 sample size.

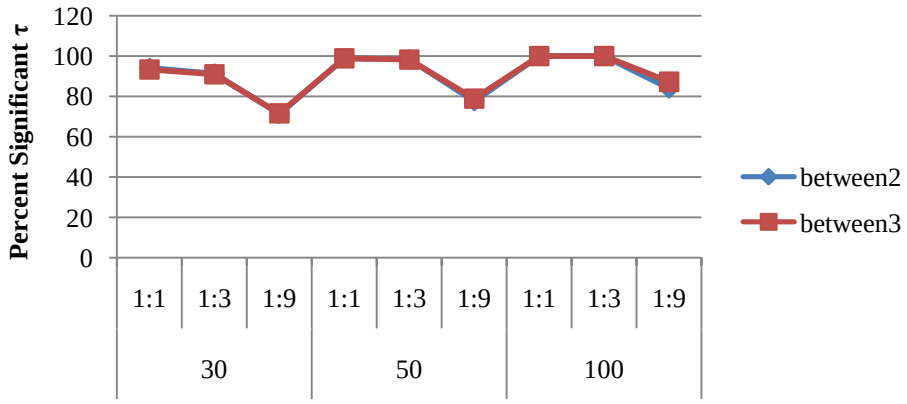
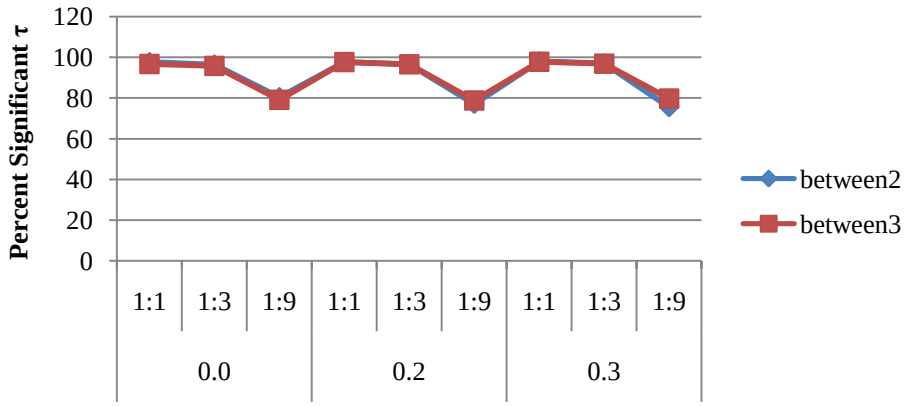


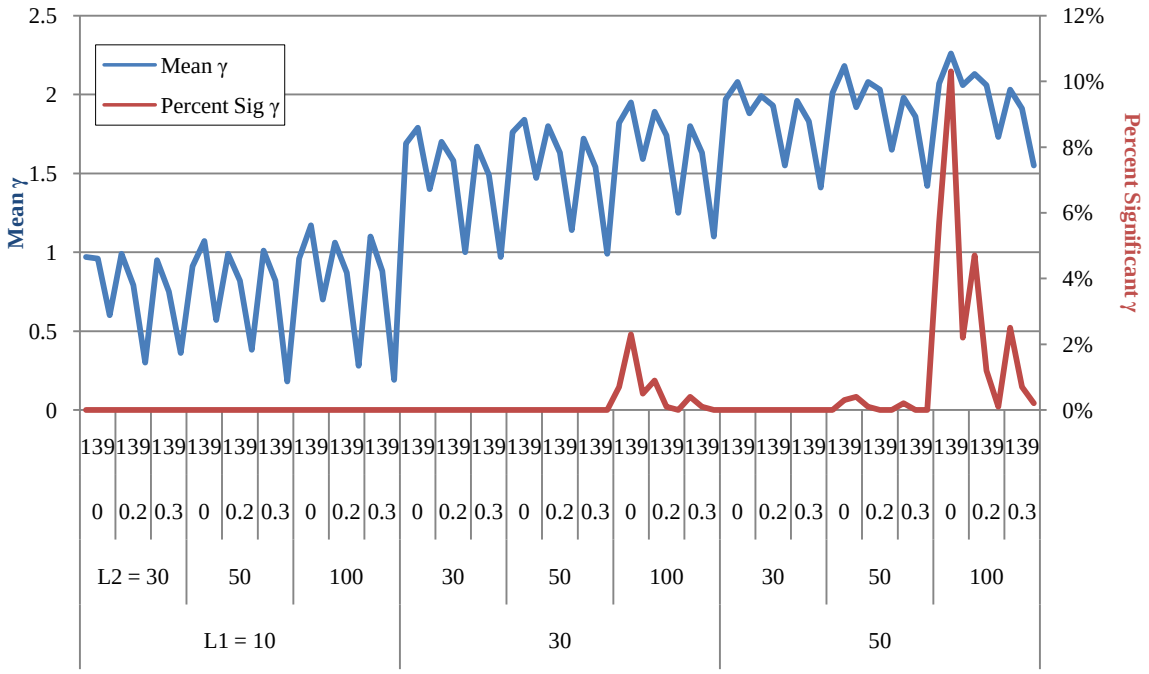
Figure 60: Mean Percentage of Significant τ_{11} per cross-level interaction and treatment-control group ratio.



Percent Significant versus Mean Value of γ_{10} and τ_{11}

Figure 61 below allows for a visual comparison of the mean values for γ_{10} and the percentage of significant γ_{10} for each simulation condition. The figure illustrates the results of the between-cluster matching using propensity scores estimated using Model 2 and Model 3. The values for mean γ_{10} remain less than 1.0 under most conditions when the sample size at level-1 is 10. The mean γ_{10} remain between 1.0 and 2.0 across conditions when level-1 sample sizes are 30 and 50, showing a consistent increase in imbalance as sample size at level-1 and level-2 increase. The values for the percent significant γ_{10}, however, show a different pattern. No significant γ_{10} values are found until sample size reaches 2,500 (50 individuals x 50 groups) after which the percent significant γ_{10} becomes more pronounced, as evidenced when sample size is equal to 3,000 (30 individuals x 100 groups) and 5,000 (50 individuals x 100 groups). This pattern suggests that the increased percent significant γ_{10} reflects the increased power to detect differences in the covariates between the groups resultant from larger sample sizes. The patterns for Model 2 are similar to those found for Model 3.

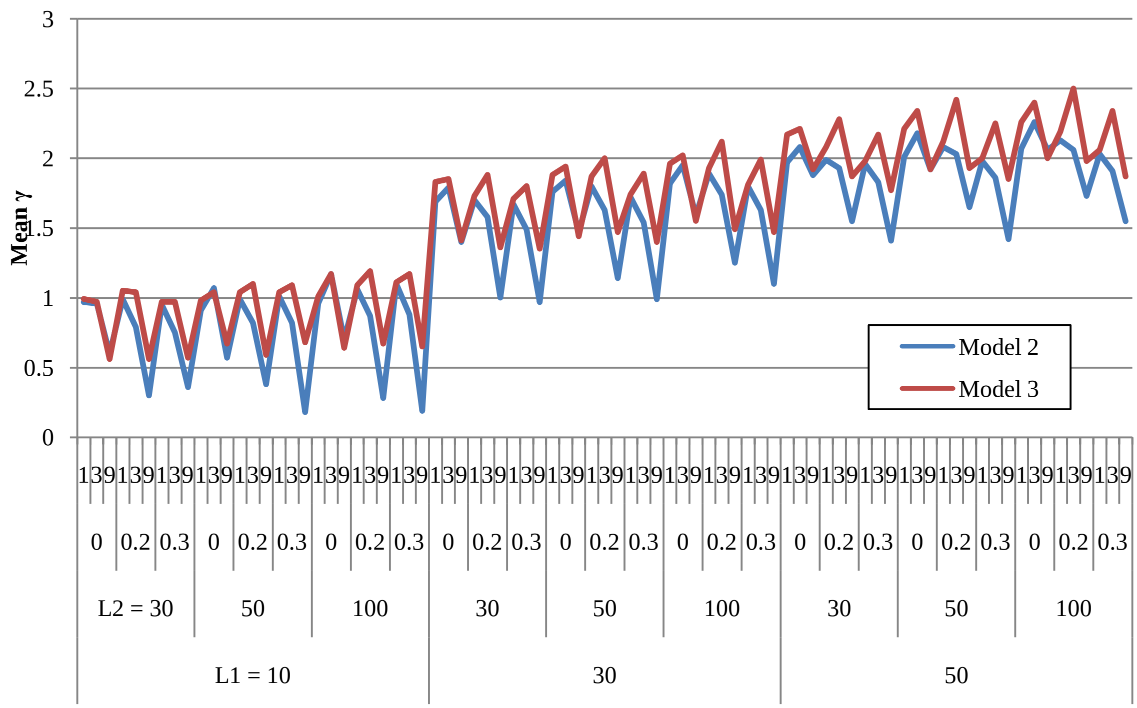
Figure 61: Overlay of Mean γ_{10} and Percent Significant γ_{10} for Between-Cluster Matching with Propensity Scores Estimated using Model 2



The mean γ_{10} across conditions is slightly higher for the propensity score estimated using Model 3 (without W) versus one that is estimated using Model 2 (with W). As the cross-level interaction increases, the difference between the performance of these two propensity score models also increases. When the cross-level interaction is 0, the mean values for γ_{10} are very similar. When the cross-level interactions are nonzero, however, the mean values for γ_{10} are only similar when the treatment-control group ratio is 1:1. In the 1:3 ratio condition, the propensity score estimated using Model 2 shows better overall balance compared to the 1:1 ratio condition, whereas the propensity score estimated using Model 3 typically shows higher mean values of γ_{10} in the 1:3 ratio condition. Both propensity score estimation models show their lowest mean values of γ_{10} in the 1:9 ratio condition. In summary, between-cluster matching methods result in smaller values for γ_{10} when matching is easiest (1:9 ratio). Additionally, a propensity

score estimated using Model 2 results in smaller values for γ_{10} than Model 3 when a nonzero cross-level interaction is present.

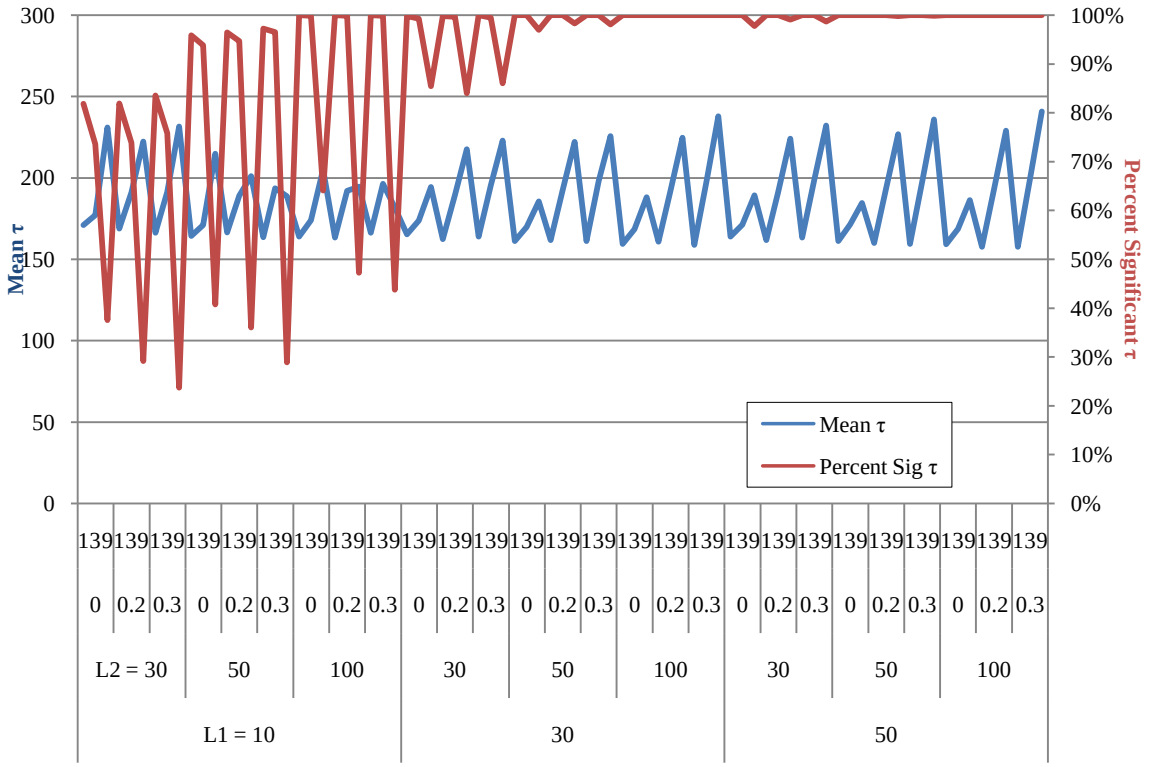
Figure 62: Mean γ_{10} for Propensity Scores Estimated using Model 2 and Model 3 applied using Between-Cluster Matching



When considering the percent significant τ_{11} and the mean values of τ_{11}, the results indicate that between-cluster matching with propensity scores estimated using Model 2 and Model 3 did not result in consistent balance across clusters. The values for percent significant τ_{11}, however, are positively related to sample size at level-1 and level-2 while the mean values of τ_{11} remained stable. As with the values of percent significant γ_{10} discussed above, the changes in the percent significant τ_{11} reflect an increase in the power to detect variance rather than increases in the variance. The influence of sample size is supported by the fact that the smallest values for percent significant τ_{11} are apparent in the 1:9 condition, when treatment group (and overall sample size) is smallest.

As described in the discussion of the mean values of γ_{10} above, the balance is better achieved across the overall sample when there are more control individuals with which to match treatment individuals (1:9 ratio); however, the variation in the balance across clusters indicated by τ_{11} is greatest in this same condition. Because matching is not limited to within clusters, these results likely indicate that the closest matches between treatment and control individuals are not occurring within each cluster. This would result in treatment group and control group members from the same clusters being matched with members from other clusters. In such a case, those individuals who remained after matching and who shared a cluster would not necessarily have similar propensity scores. Additionally, some clusters may contain only treatment group members while others contain only control group members after application of between-cluster matching.

Figure 63: Overlay of Mean τ_{11} and Percent Significant τ_{11} for Between-Cluster Matching with Propensity Scores Estimated using Model 2

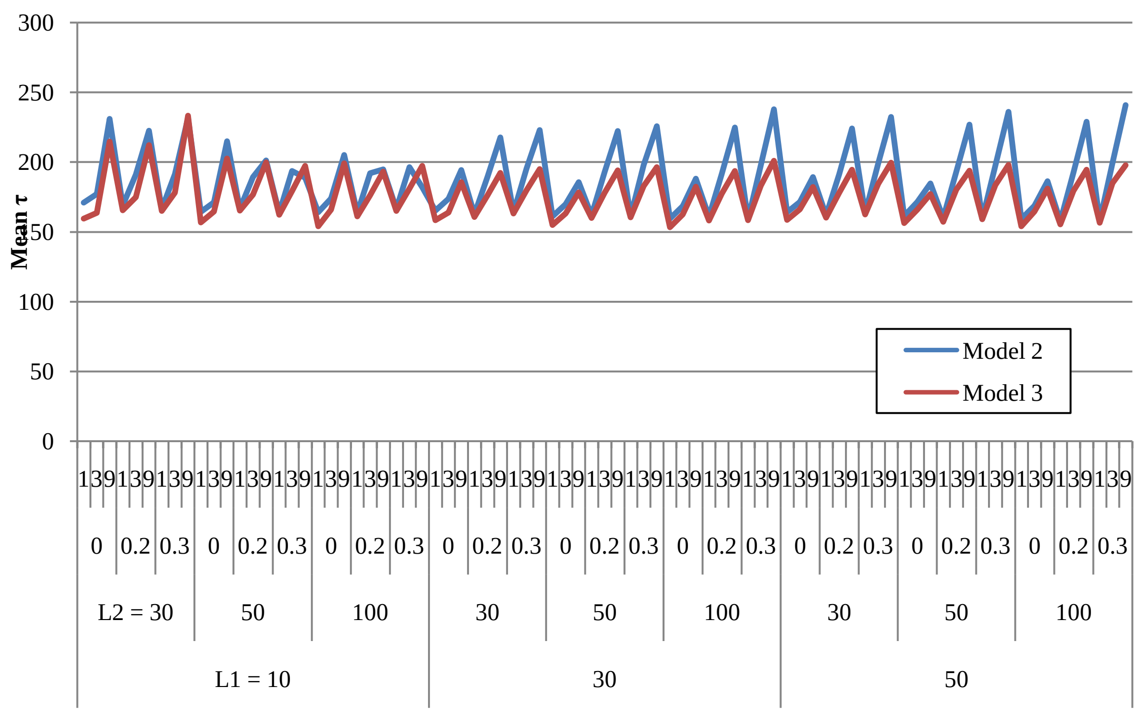


When comparing the mean τ_{11} between propensity scores estimated using Model 2 to Model 3, several differences in the results are apparent. As the cross-level interaction increases, the propensity scores that include W showed slightly higher mean values of τ_{11} (e.g., evidenced greater variance in balance across clusters) than was shown for Model 3, especially with larger sample sizes. This difference between propensity score performance was greatest when cross-level interactions were larger and when the number of treatment group members was smallest (e.g., 1:9 ratio). A possible explanation of this effect is that when no cross-level interaction is present, adding the W to the propensity score estimation model does not affect the variance in the propensity score; however, as the cross level interaction increases, adding the W to the equation can actually inflate the

variance. This would result in less precise propensity scores and subsequently poorer matches.

Another possible explanation for this finding apparent in the 1:9 ratio condition follows. Adding W to the propensity score estimation model results in more matches outside of the cluster than when using a propensity score estimated without W. The overall balance was better for Model 2 than for Model 3 in the 1:9 condition, which suggests that propensity scores were efficiently estimated with Ws included even when treatment group members within the cluster were few. These findings together suggest that better matches occur outside the cluster, and that better matches are accomplished using Model 2 than when using Model 3.

Figure 64: Mean τ_{11} from Between-Cluster Matching with Propensity Scores Estimated using Model 2 and Model 3



Balance Achievement in Predictor Covariates: The Standardized Mean Difference

Sample Size at Level-1 and Level-2. The SMD between level-1 covariates across the sample as a whole was below 10 in most cells. The relationship between sample-size and SMD was negative: As the sample-size increases at level-1 or level-2, the SMD decreases (see Table 40).

Table 40:

SMD for Covariates Resultant from Between-Cluster Matching using Propensity Scores Estimated using Model 2 and Model 3

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
Model 2										
30	6.4	3.7	2.9	11.7	6.8	5.2	10.5	5.8	4.5	6.4
50	4.7	2.9	2.4	8.8	5.2	4.1	7.6	4.3	3.4	4.8
100	3.3	2.2	2.0	6.0	3.7	3.0	5.2	3.1	2.5	3.4
Mean	4.8	2.9	2.4	8.9	5.2	4.1	7.7	4.4	3.5	4.9
Model 3										
30	5.8	3.2	2.5	11.6	6.6	5.1	9.9	5.5	4.3	6.0
50	4.3	2.4	1.9	8.8	5.0	4.0	7.5	4.2	3.3	4.6
100	2.8	1.6	1.3	6.0	3.5	2.8	5.0	2.9	2.3	3.1
Mean	4.3	2.4	1.9	8.8	5.0	3.9	7.5	4.2	3.3	4.6

Cross-Level Interaction. When applying a propensity score estimated using Model 2 to balance covariates with between-cluster matching, the relationship between

the strength of the cross-level interaction and the SMD is 0. A small negative relationship exists, however, when the propensity score is estimated using Model 3: As the cross-level interaction increases, the SMD decreases. This finding was consistent within each sample-size condition and is presented in Table 41.

Table 41:

SMD for X_1 per Cross-Level Interaction Resultant from Between-Cluster Matching using Regression Propensity Score Estimation Models per Level-1 and Level-2 Sample Sizes

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	6.7	4.8	3.2	3.7	2.8	1.9	2.8	2.1	1.5	3.3
0.2	6.3	4.7	3.4	3.6	3.0	2.4	3.0	2.5	2.1	3.4
0.3	6.2	4.6	3.3	3.6	2.9	2.4	3.0	2.7	2.3	3.4
	Model 3									
0.0	6.2	4.7	3.1	3.5	2.6	1.8	2.7	2.0	1.5	3.1
0.2	5.5	4.3	2.7	3.1	2.3	1.6	2.4	1.8	1.2	2.8
0.3	5.5	4.0	2.6	3.0	2.2	1.5	2.3	1.7	1.2	2.7

Treatment-Control Group Ratio. The 1:9 ratio condition consistently showed the highest SMD, and the 1:1 condition consistently showed the smallest SMD. Results suggest that no interaction effects exist between this relationship and sample-size conditions or the cross-level interaction conditions.

Figure 65: Mean SMD for X_1 per Level-1 Sample Size

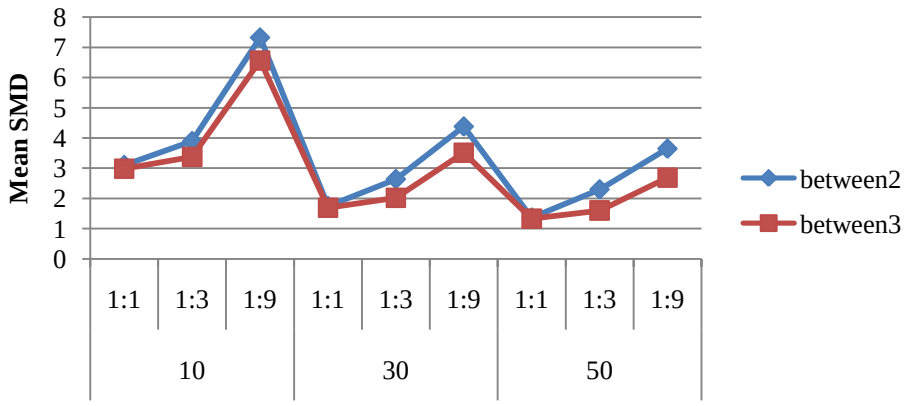


Figure 66: Mean SMD for X_1 per Level-2 Sample Size

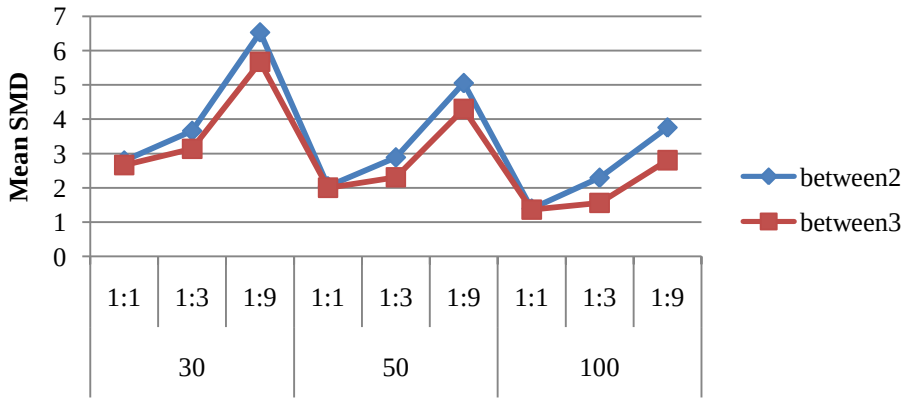


Figure 67: Mean SMD for X_1 per Cross-Level Interaction

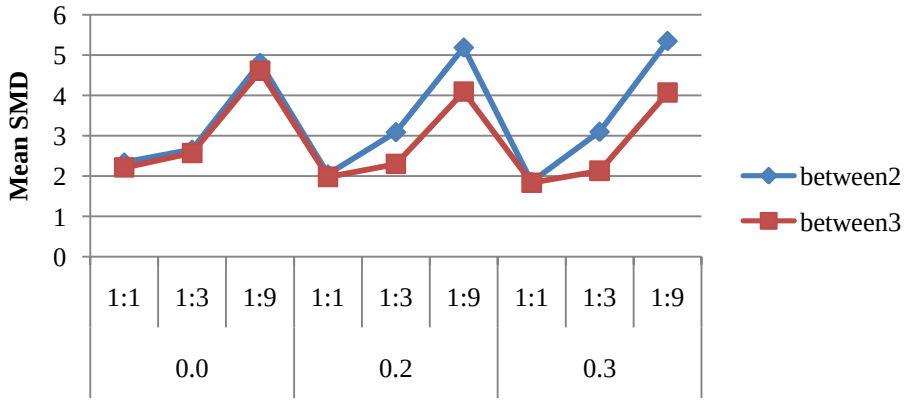
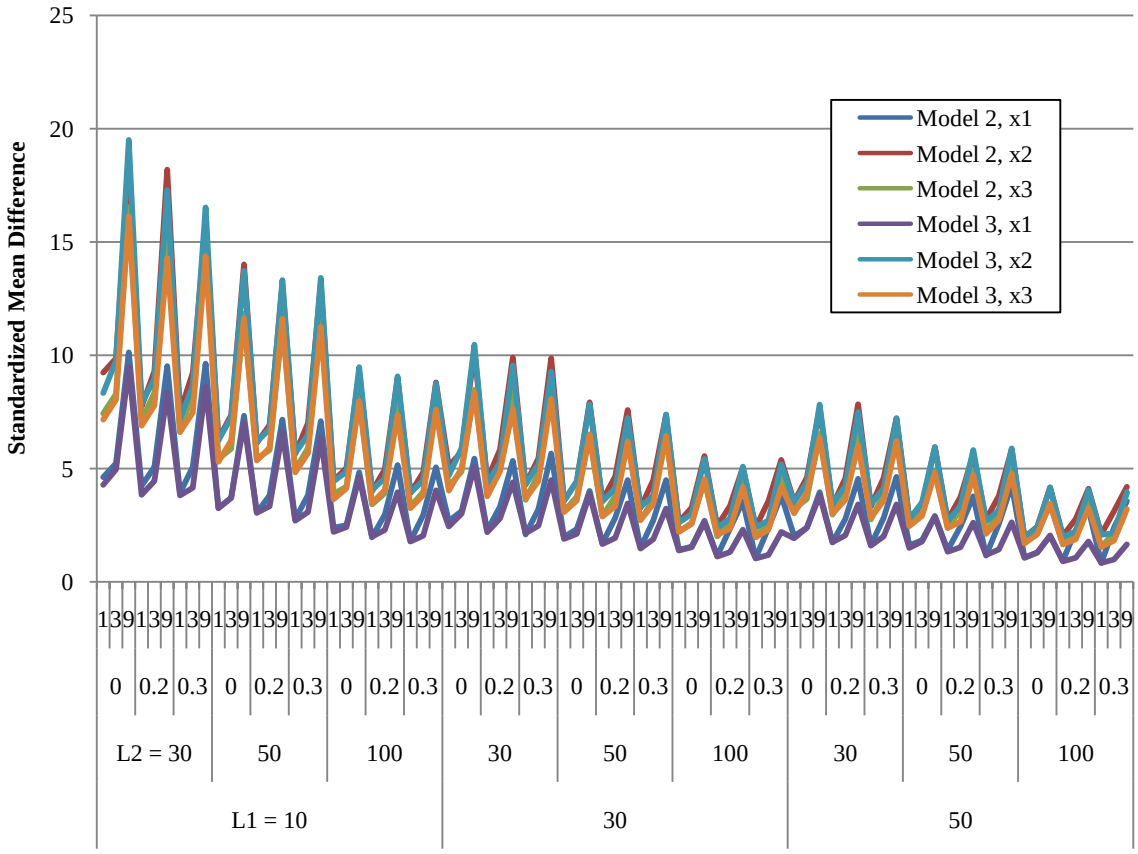


Figure 68: Standardized mean Differences for X_1, X_2, and X_3 per Simulation Condition for Between-Cluster Matching with Propensity Scores Estimated using Model 2 and Model 3



SMD versus Mean γ_{10}

The patterns apparent in the SMD and the mean γ_{10} for between-cluster matching are nearly opposite reflections of one another. This quality is clearly apparent in results from application of propensity scores estimated from both Models 2 and 3 (see Figure 69 and Figure 70, respectively). The explanation for these opposite patterns is likely related to the calculation of these measures. The SMD does not take into account the effects of clustering, whereas the mean γ_{10} does take clustering into account. If the clustering is not taken into account, then the variance in the balance between groups will be attributed to

differences in the groups rather than the group membership. Greater variance in the scores of treatment and control groups would subsequently result in lower SMDs. This explanation is supported by the figures below, especially Figure 70 which illustrates results from Model 3, which does not include a cluster-level predictor in the estimation. In this figure, it is apparent that as the cross-level interaction increases, the SMD decreases. In this case, the influence of W is not included in the estimation of the propensity score, which results in greater variance between treatment and control group members' propensity scores, and subsequently smaller values of the SMD.

Figure 69: Overlay of SMD and Mean γ_{10} for Between-Cluster Matching with Propensity Scores Estimated using Logistic Regression with a Cluster-Level Predictor for X_1

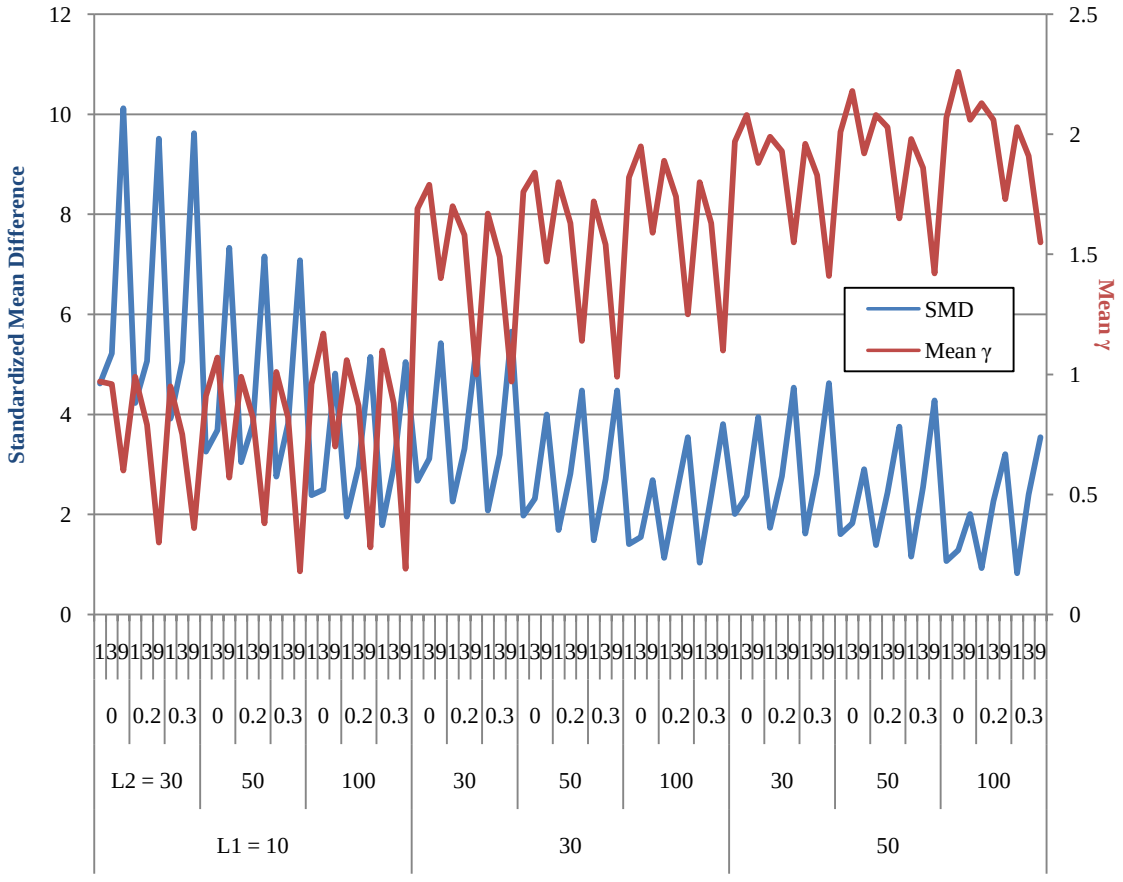
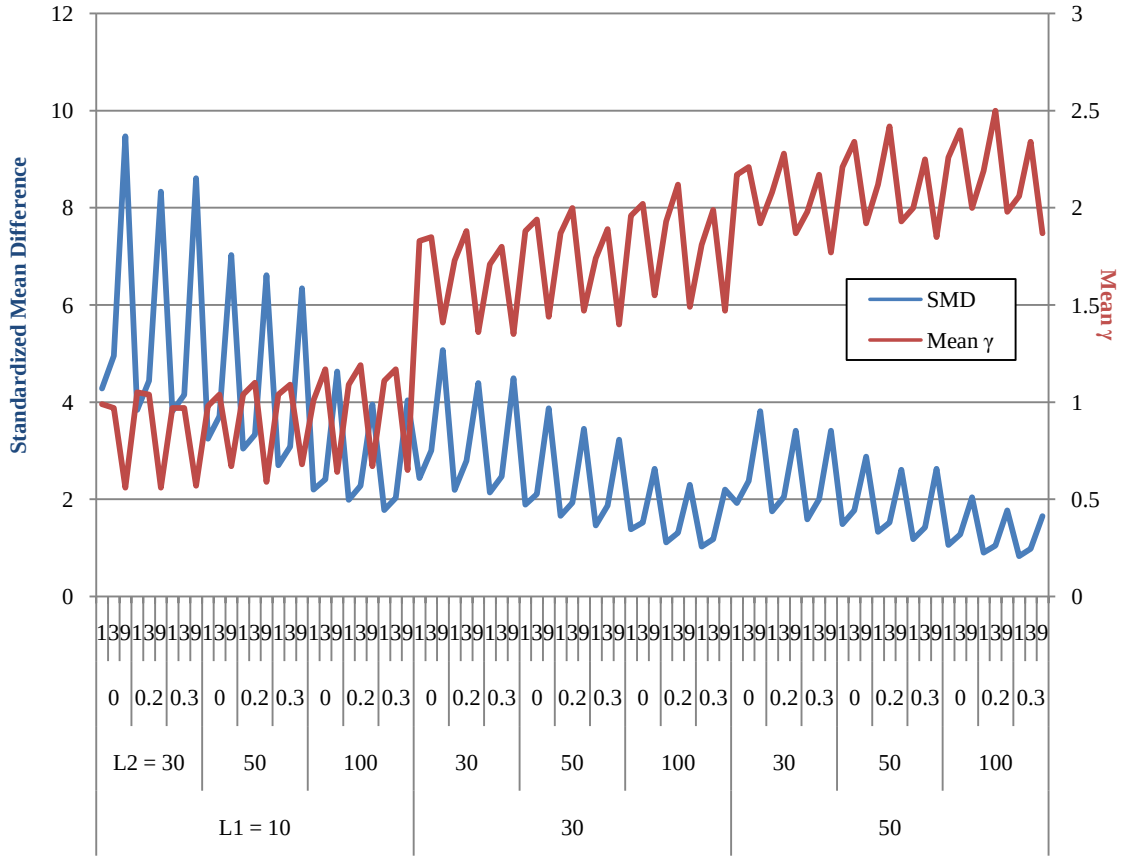


Figure 70: Overlay of SMD and Mean γ_{10} for Between-Cluster Matching with Propensity Scores Estimated using Logistic Regression without a Cluster-Level Predictor



Balance Achievement in Predictor Covariates: Percentage Significant γ_{10}

Sample Size at Level-1 and Level-2. When applying a propensity score that is estimated using Models 2 or Model 3 to balance covariates through quintile stratification, a positive relationship is evident between the percent significant γ_{10} and sample size conditions at both level-1 and level-2: As sample size increases, the percentage significant γ_{10} increases. This pattern is evident in the balance achieved in X_1, only, which is the covariate with the strongest correlation with the treatment assignment, Z. The percent significant γ_{10} in X_1 that results from using a propensity score estimated using Model 2 was approximately twice the percentage resulting from a propensity score

estimated using Model 3. The results for these two models per sample-size condition are presented in Table 42.

Table 42:

Average Percentage of Significant γ_{10} for Covariates Resultant from Quintile Stratification using Propensity Scores Estimated using Model 2 and Model 3

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
Model 2										
30	0.6	2.3	3.9	0.0	0.1	0.1	0.0	0.0	0.0	0.8
50	1.3	4.8	7.3	0.0	0.1	0.1	0.0	0.0	0.0	1.5
100	3.8	15.5	21.9	0.0	0.2	0.5	0.0	0.0	0.0	4.7
Mean	1.9	7.5	11.0	0.0	0.1	0.2	0.0	0.0	0.0	2.3
Model 3										
30	0.5	1.2	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.4
50	0.7	2.3	3.7	0.0	0.0	0.0	0.0	0.0	0.0	0.7
100	2.2	8.2	12.3	0.0	0.0	0.1	0.0	0.0	0.0	2.5
Mean	1.1	3.9	6.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2

Cross-Level Interaction. The relationship of the cross-level interaction and the percentage significant γ_{10} is different per propensity score estimation model. When the propensity score is estimated including the cluster-level predictor, there is essentially no relationship. When the propensity score is estimated without the cluster-level predictor,

the relationship is negative: As the strength of the cross-level interaction increases, the percentage significant γ_{10} decreases. These patterns are consistent across sample-size conditions, as illustrated in Table 43.

Table 43:

Average Percentage of Significant γ_{10} for X_1 per Cross-Level Interaction Resultant from Quintile Stratification using Propensity Scores Estimated using Model 2 and Model 3

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	0.3	1.2	3.6	2.1	3.9	15.0	3.4	6.7	21.5	6.4
0.2	0.7	1.2	4.5	2.3	5.4	16.2	4.0	7.7	22.5	7.2
0.3	0.9	1.3	3.3	2.4	5.3	15.4	4.3	7.6	21.6	6.9
	Model 3									
0.0	0.5	1.1	3.8	2.1	3.5	15.8	3.3	6.6	23.8	6.7
0.2	0.6	0.5	1.4	0.8	1.9	5.2	1.5	2.2	7.6	2.4
0.3	0.3	0.4	1.3	0.6	1.6	3.6	1.3	2.2	5.5	1.9

Treatment-Control Group Ratio. The relationship of the treatment-control group ratio and the percent significant γ_{10} is positive for both propensity score estimation Model 2 and Model 3: The percent significant γ_{10} is smallest when the ratio is 1:1, larger at 1:3, and largest at 1:9. The percent significant γ_{10} under the 1:1 and 1:3 ratio conditions are much closer in value to each other than either is to the 1:9 ratio condition; however, the

propensity score that is estimated using Model 2 results in a consistently higher percent significant γ_{10} in the 1:9 ratio condition than that estimated using Model 3. This pattern is apparent across level-1 sample size conditions, level-2 sample-size conditions, and cross-level interaction conditions with one exception: When there is no cross-level interaction, the patterns per propensity score estimation method are approximately identical. These relationships are illustrated in Figure 71, Figure 72, and Figure 73.

Figure 71: Mean percent significant γ_{10} per treatment-control ratio and level-1 sample size.

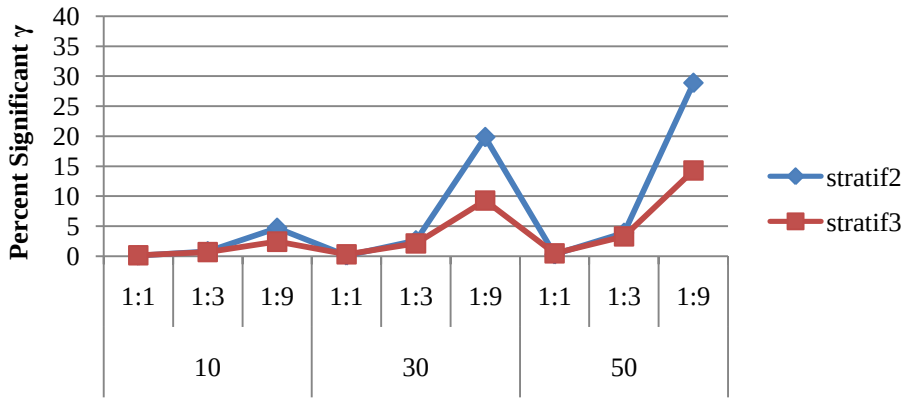


Figure 72: Mean percent significant γ_{10} per treatment-control ratio and level-2 sample size.

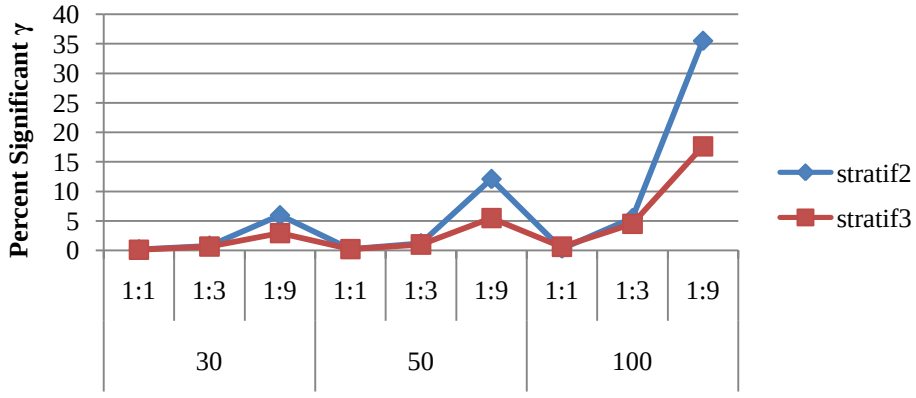


Figure 73: Mean percent significant γ_{10} per cross-level interaction and treatment-control ratio

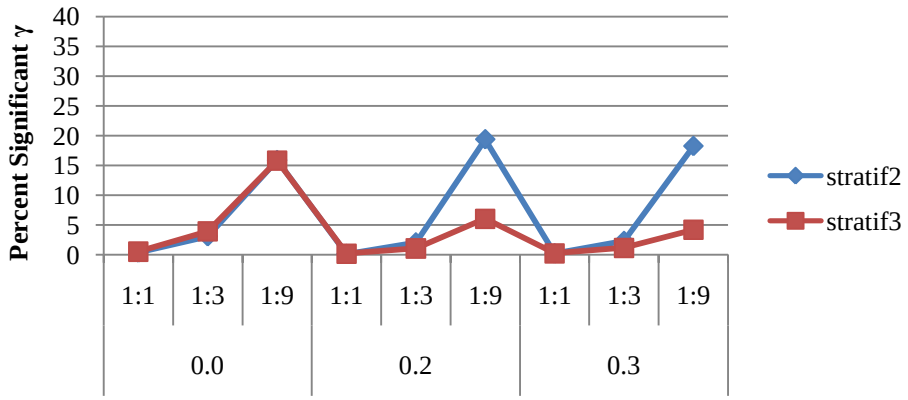
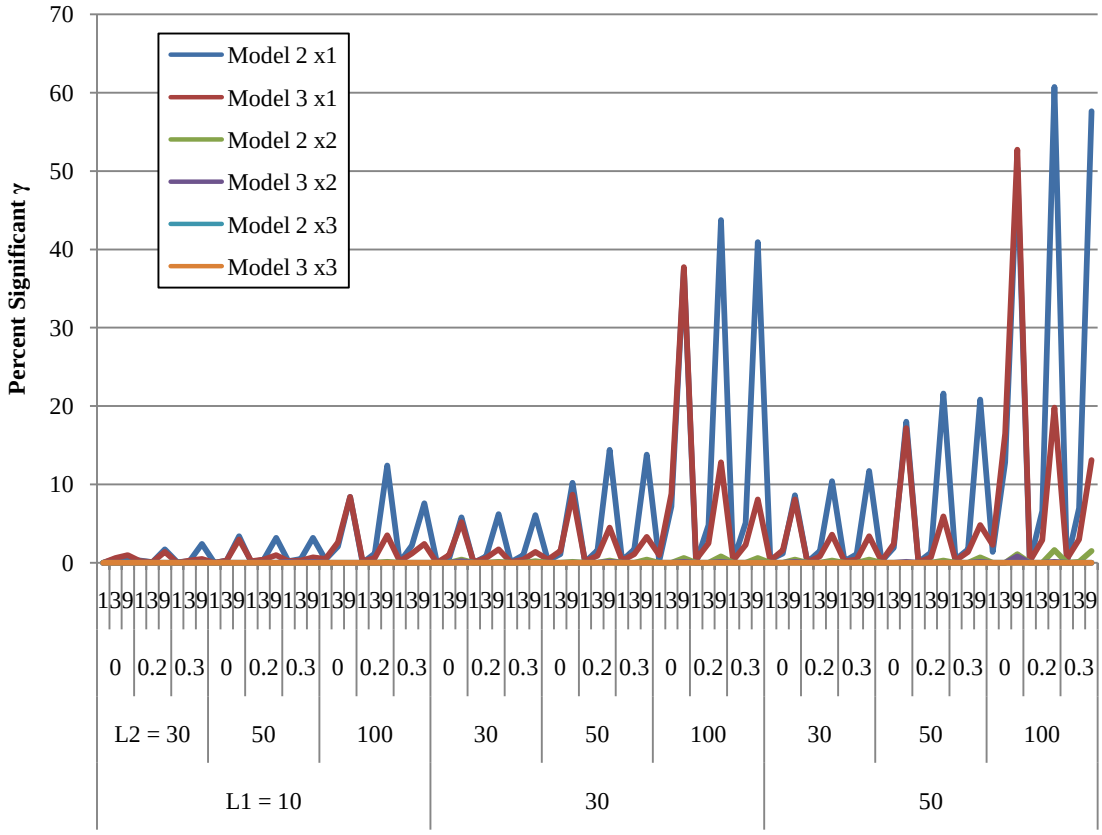


Figure 74: Percent Significant γ_{10} for X_1, X_2, and X_3 across Simulation Conditions for Quintile Stratification using Propensity Scores Estimated using Model 2 and Model 3



Variation in Balance Achievement within Clusters: Percentage Significant τ_{11}

Sample Size at Level-1 and Level-2. The percent significant τ_{11} shows a positive relationship with sample-size when applying propensity scores estimated using Models 2 and 3 and applied through quintile stratification. Overall, the values for τ_{11} were significant in almost 100% of replications within each sample-size condition. Results are presented in Table 44.

Table 44:

Average Percentage of Significant τ_{11} Resultant from Quintile Stratification using Propensity Scores Estimated using Model 2 and Model 3

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
					Model 2					
30	73.0	98.0	99.7	76.1	98.9	99.9	83.7	99.7	100.0	92.1
50	89.3	99.8	100.0	90.0	100.0	100.0	94.1	100.0	100.0	97.0
100	98.4	100.0	100.0	97.9	100.0	100.0	99.5	100.0	100.0	99.5
Mean	86.9	99.3	99.9	88.0	99.6	100.0	92.4	99.9	100.0	96.2
					Model 3					
30	73.0	98.0	99.7	76.7	98.8	99.9	85.4	99.7	100.0	92.4
50	88.8	99.8	100.0	90.3	99.9	100.0	94.5	100.0	100.0	97.0
100	98.0	100.0	100.0	97.9	100.0	100.0	99.5	100.0	100.0	99.5
Mean	86.6	99.3	99.9	88.3	99.6	100.0	93.1	99.9	100.0	96.3

Cross-Level Interaction. The strength of the cross-level interaction showed a negative relationship with percent of significant τ_{11}: As the cross-level interaction increases, the percent of significant τ_{11} decreases. These patterns are nearly identical per propensity score estimation model. The results for X_1 are presented in Table 45.

Table 45:

Average Percentage of Significant τ_{11} for X_1 per Cross-Level Interaction Resultant from Quintile Stratification using Propensity Scores Estimated using Model 2 and Model 3

$\rho_{(wx)z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	80.7	93.7	99.7	99.5	100.0	100.0	99.9	100.0	100.0	97.1
0.2	72.4	90.5	98.7	98.7	100.0	100.0	99.8	100.0	100.0	95.6
0.3	65.8	83.7	96.9	95.8	99.4	100.0	99.4	100.0	100.0	93.5
	Model 3									
0.0	79.7	94.0	99.7	99.7	100.0	100.0	100.0	100.0	100.0	97.0
0.2	73.6	89.4	98.3	98.6	99.9	100.0	99.8	100.0	100.0	95.5
0.3	65.7	83.2	96.0	95.8	99.5	99.9	99.4	100.0	100.0	93.3

Treatment-Control Group Ratio. The percent of significant τ_{11} shows a negative relationship with the treatment-control group ratio: as the difference in the ratio increases, the percent of significant τ_{11} decreases. This relationship was consistent across sample-size conditions and the strength of the cross-level interactions. This relationship is not

apparent once sample-size at level-1 reaches 30 due to the fact that approximately 100% of the replications in cells had a significant τ_{11}. These relationships are presented in Figure 75, Figure 76, and Figure77.

Figure 75: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-1 sample size.

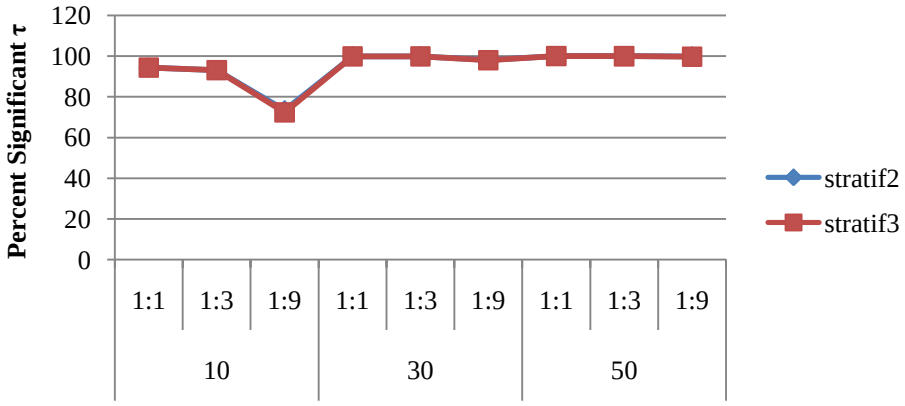


Figure 76: Mean Percentage of Significant τ_{11} per treatment-control ratio and level-2 sample size.

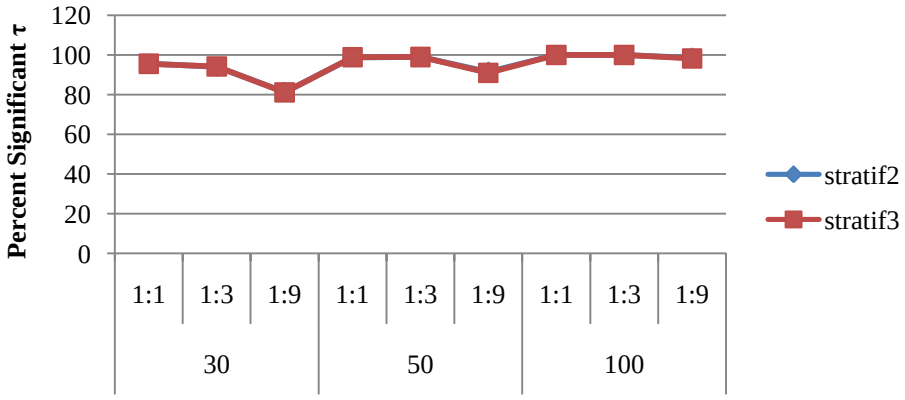
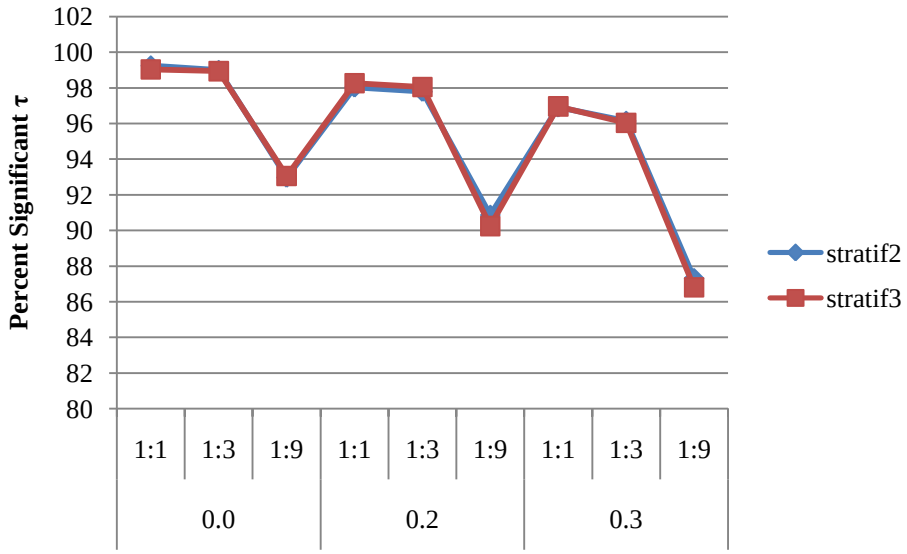


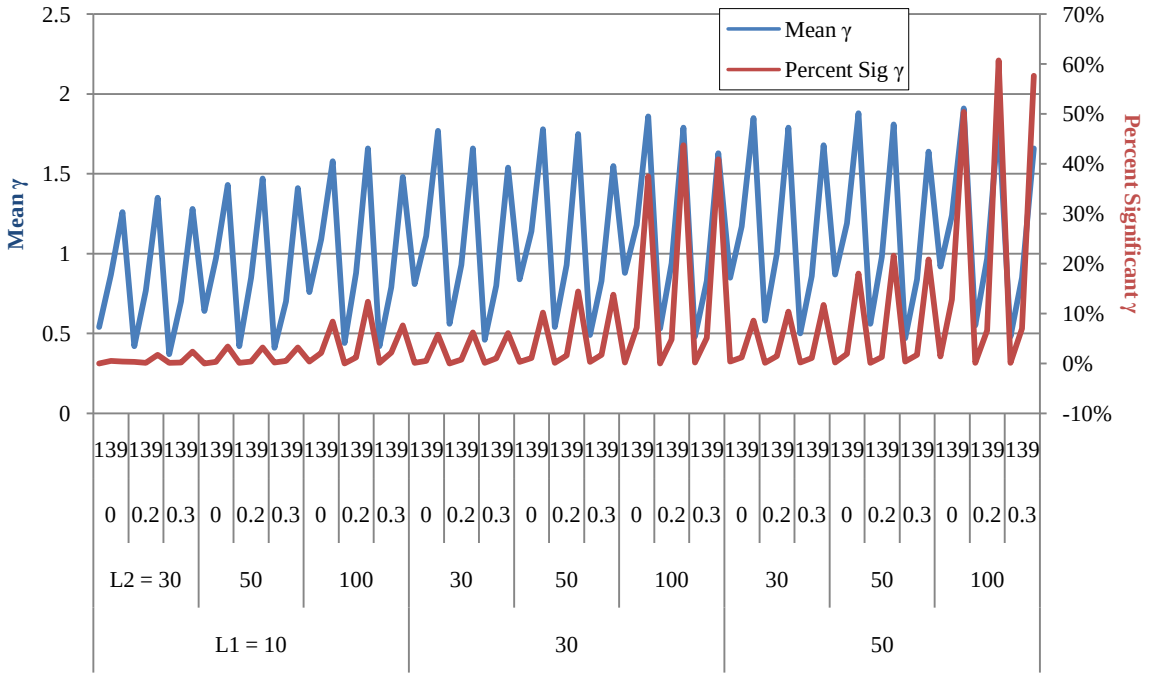
Figure 77: Mean Percentage of Significant τ_{11} per cross-level interaction and treatment-control group ratio.



Percent Significant versus Mean Value of γ_{10} and τ_{11}

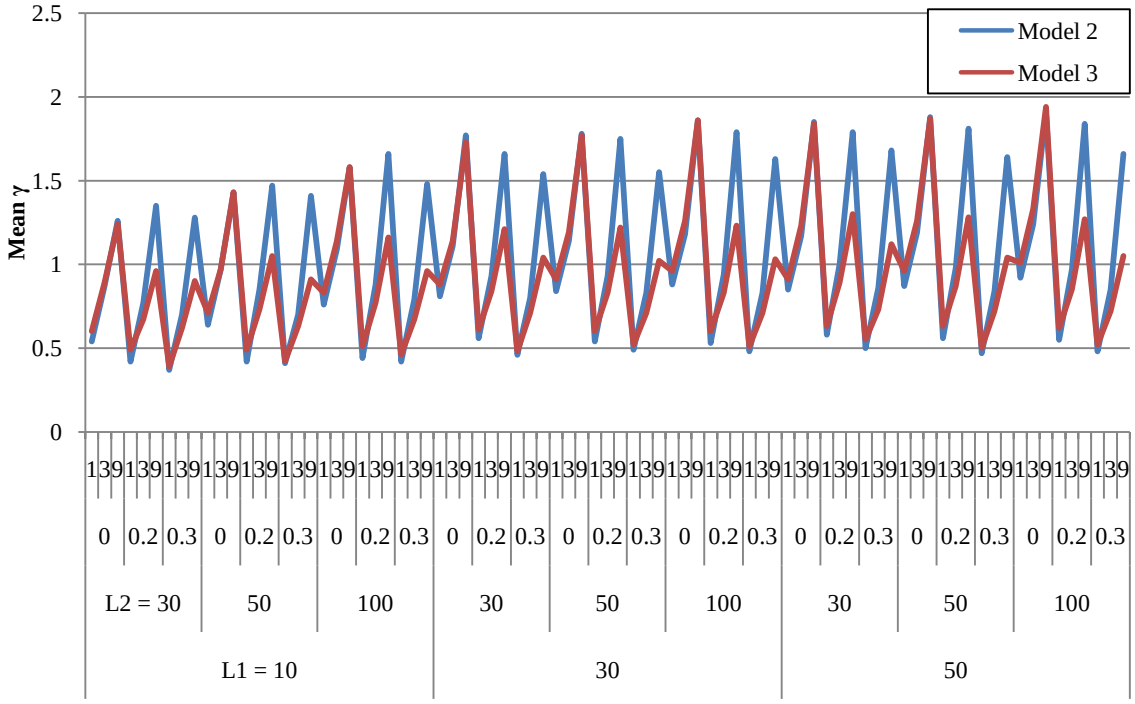
When examining the results of quintile stratification when using a propensity score estimated using Models 2 and 3, both the mean γ_{10} and percent significant γ_{10} show similar patterns across simulation conditions; however, when the sample-size at level-2 is 100, the percent significant γ_{10} spikes whereas the mean γ_{10} remains fairly constant. This spike in percent significant γ_{10} is likely due to the power to detect differences due to large sample sizes. Both indicators of covariate balance provide support to the conclusion that propensity scores estimated using Model 2 and Model 3, when applied using quintile stratification, result in better covariate balance when the treatment-control group ratio is 1:1 as compared to results in the 1:3 and 1:9 ratio conditions.

Figure 78: Overlay of Mean γ_{10} and Percent Significant γ_{10} for Quintile Stratification Matching with Propensity Scores Estimated using Model 2



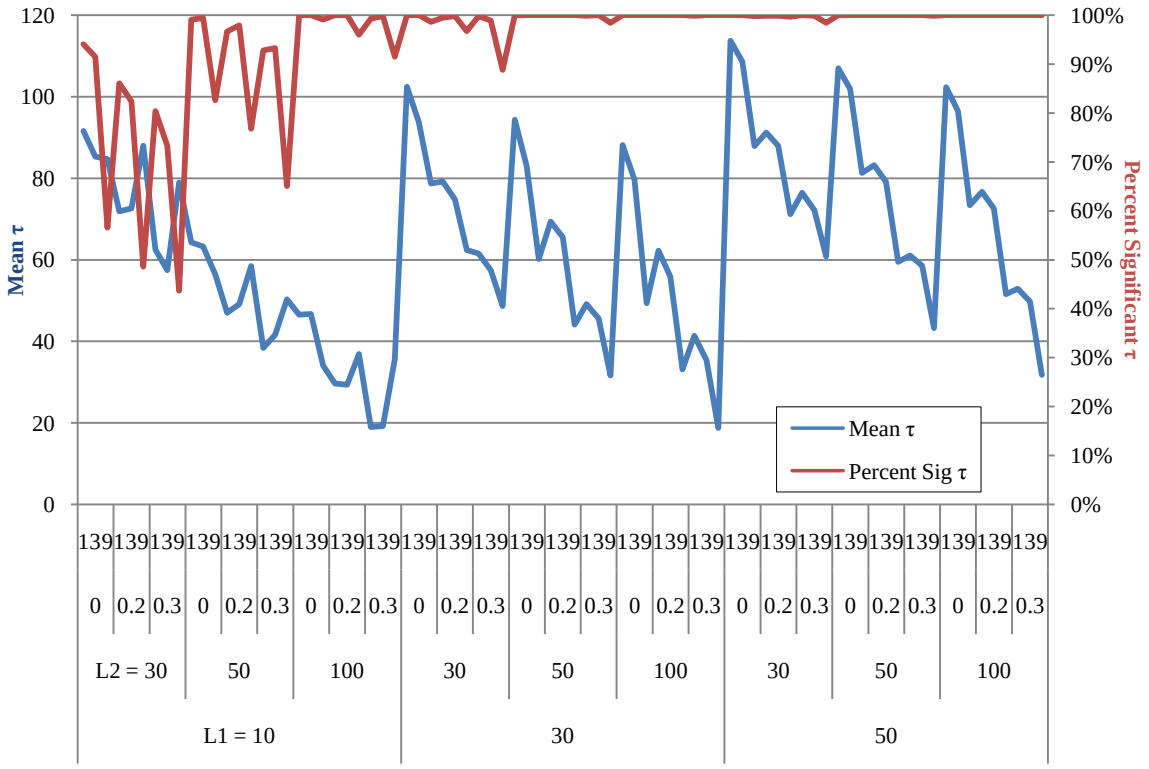
The overall balance that results from using a propensity score using Model 2 showed almost identical results to a propensity score estimated using Model 3 when the cross-level interaction was 0 and when treatment-control group ratios were 1:1. When the cross-level interactions were larger and the treatment-control group ratio was either 1:3 or 1:9, however, Model 3 showed smaller values for mean γ_{10} than did Model 2. This finding is likely resultant from the challenge of estimating cluster-level effects using Model 2 when the number of treatment group members per cluster is small.

Figure 79: Mean γ_{10} from Quintile Stratification using Propensity Scores Estimated using Model 2 and Model 3



The influence of sample size is apparent in the percent significant τ_{11}. Once sample size at level-1 reached 30, nearly all replications per cell showed significant variance in the balance across clusters. A possible explanation for this finding is related to the increased power to detect differences in slopes across clusters due to larger sample sizes rather than the values of τ_{11}. The values for mean τ_{11} show a clear negative relationship with the cross-level interaction, with level-2 sample size, and with the treatment-control group ratio (smaller values of mean τ_{11} in the 1:9 ratio condition versus the 1:1 ratio condition). As sample-size at level-1 increases, however, the values for mean τ_{11} also increase. The values for mean τ_{11} resulting from Model 2 and from Model 3 were almost identical across conditions.

Figure 80: Overlay of Mean τ_{11} and Percent Significant τ_{11} for Quintile Stratification with Propensity Scores Estimated using Model 2



Balance Achievement in Predictor Covariates: The Standardized Mean Difference

Sample Size at Level-1 and Level-2. Findings suggest that a negative relationship exists between the SMD and the sample-size at level-1 and level-2 when quintile stratification is used: As sample-size increases, the SMD decreases. The average SMD resulting from quintile stratification using a propensity score estimated with Model 2 was consistently higher by a few percentage points than those resultant from propensity scores estimated using Model 3. These findings are presented in Table 46.

Table 46:

SMD Resultant from Quintile Stratification with Propensity Score Estimated using Model 2 and Model 3

Level-2 Sample Size	X_1			X_2			X_3			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
					Model 2					
30	33.0	23.4	20.2	32.9	23.5	20.3	31.2	21.0	17.3	24.8
50	27.5	19.1	17.2	27.7	19.6	17.5	25.0	16.1	13.5	20.4
100	20.9	15.6	14.5	21.7	16.2	14.8	18.2	11.7	9.9	15.9
Mean	27.1	19.4	17.3	27.4	19.8	17.5	24.8	16.3	13.6	20.4
					Model 3					
30	31.3	21.1	17.7	30.8	20.6	17.3	30.3	20.1	16.5	22.9
50	25.3	16.7	14.3	24.8	16.5	13.9	24.4	15.5	12.9	18.3
100	18.7	12.8	11.2	18.5	12.4	10.7	17.7	11.3	9.4	13.6
Mean	25.1	16.9	14.4	24.7	16.5	14.0	24.2	15.7	12.9	18.3

Cross-Level Interaction. The relationship of the cross-level interaction and the SMD is different per propensity score estimation model. When the propensity score is estimated using Model 2, this relationship is trivial. When the propensity score is estimated using Model 3, the relationship is negative, although small: As the strength of the cross-level interaction increases, the percentage significant γ_{10} decreases. These patterns are consistent across sample-size conditions, as illustrated in Table 47.

Table 47:

SMD for X_1 per Cross-Level Interaction Resultant from Quintile Stratification using a Multilevel Propensity Score Estimation Model per Level-1 and Level-2 Sample Sizes

$\rho_{(WX)Z}$	Level-1 Sample Size									Mean
	10			30			50			
	Level-2 Sample Size			Level-2 Sample Size			Level-2 Sample Size			
	30	50	100	30	50	100	30	50	100	
	Model 2									
0.0	32.3	26.9	20.5	23.0	18.4	14.4	19.6	16.1	12.6	20.4
0.2	33.4	27.9	21.0	23.5	19.4	16.2	20.6	17.6	15.3	21.7
0.3	33.4	27.9	21.2	23.7	19.7	16.3	20.6	18.0	15.5	21.8
	Model 3									
0.0	32.1	26.3	20.2	22.7	18.5	14.4	19.3	16.0	12.9	20.3
0.2	31.1	25.3	18.5	20.9	16.4	12.5	17.6	13.9	10.9	18.6
0.3	30.6	24.4	17.4	19.9	15.3	11.6	16.2	12.9	9.8	17.6

Treatment-Control Group Ratio. A positive relationship is apparent between the mean SMD and the treatment-control ratio: as the difference in the number of treatment and control group members increases, the mean SMD increases. This pattern is consistent across sample-size conditions and cross-level interaction conditions as illustrated in Figure 81, Figure 82, and Figure 83.

Figure 81: Mean SMD for X_1 per Level-1 Sample Size

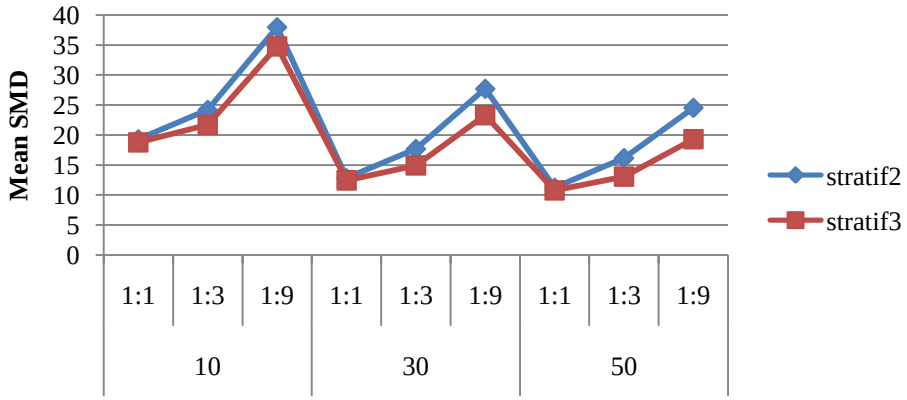


Figure 82: Mean SMD for X_1 per Level-2 Sample Size

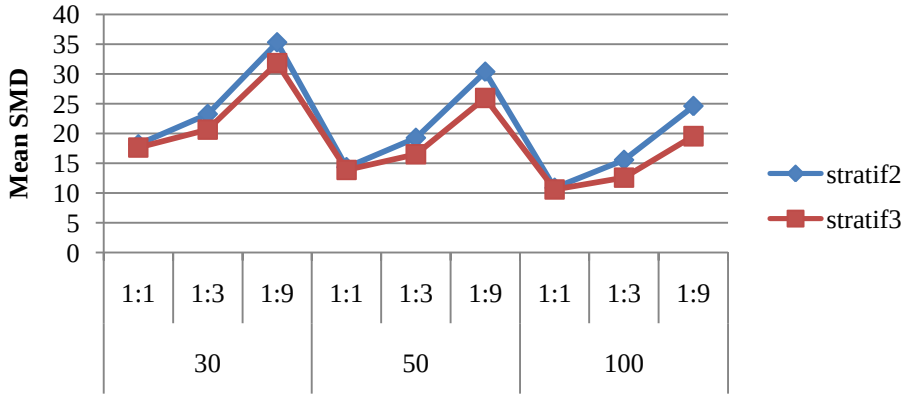
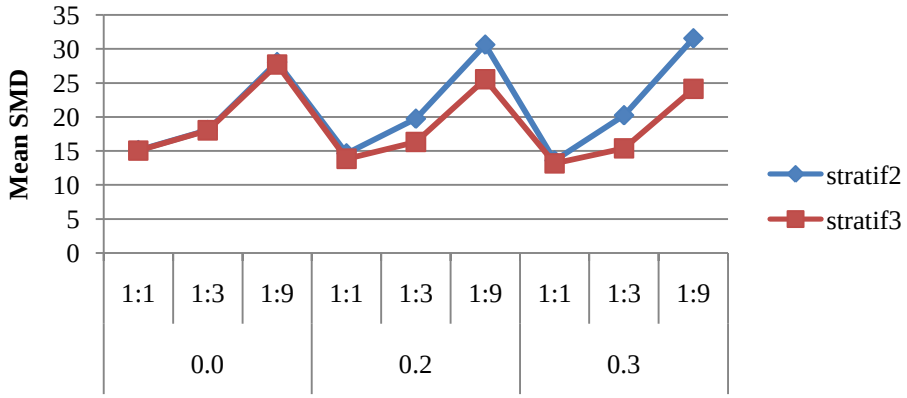


Figure 83: Mean SMD for X_1 per Cross-Level Interaction



Focusing on the average SMD across conditions obscures one particular relationship that is suggested in the figures above but made clear in Figure 84 below when the value for the SMD for every simulation condition is presented. As previously noted, for both Model 2 and Model 3, the SMDs are practically identical when the cross-level interaction is 0; however, as the cross-level interaction increases, the relationship between SMD and treatment-control group ratio differ. When using a propensity score that is estimated using Model 2, the 1:9 treatment-control ratio is associated with larger values of SMD. When W is not included in the propensity score estimation (Model 3), this relationship between SMD and treatment-control ratio is less pronounced. Additionally, when Model 2 is used, the cross-level interaction is positively related to the SMD; whereas, with Model 3, the cross-level interaction is negatively related to SMD. This difference is likely related to the challenge of estimating the effects of W in Model 2 when the number of treatment group members per cluster is small (e.g., 1:9 condition).

Figure 84: Mean SMD for X_1 per Propensity Score Estimation Method across Simulation Conditions

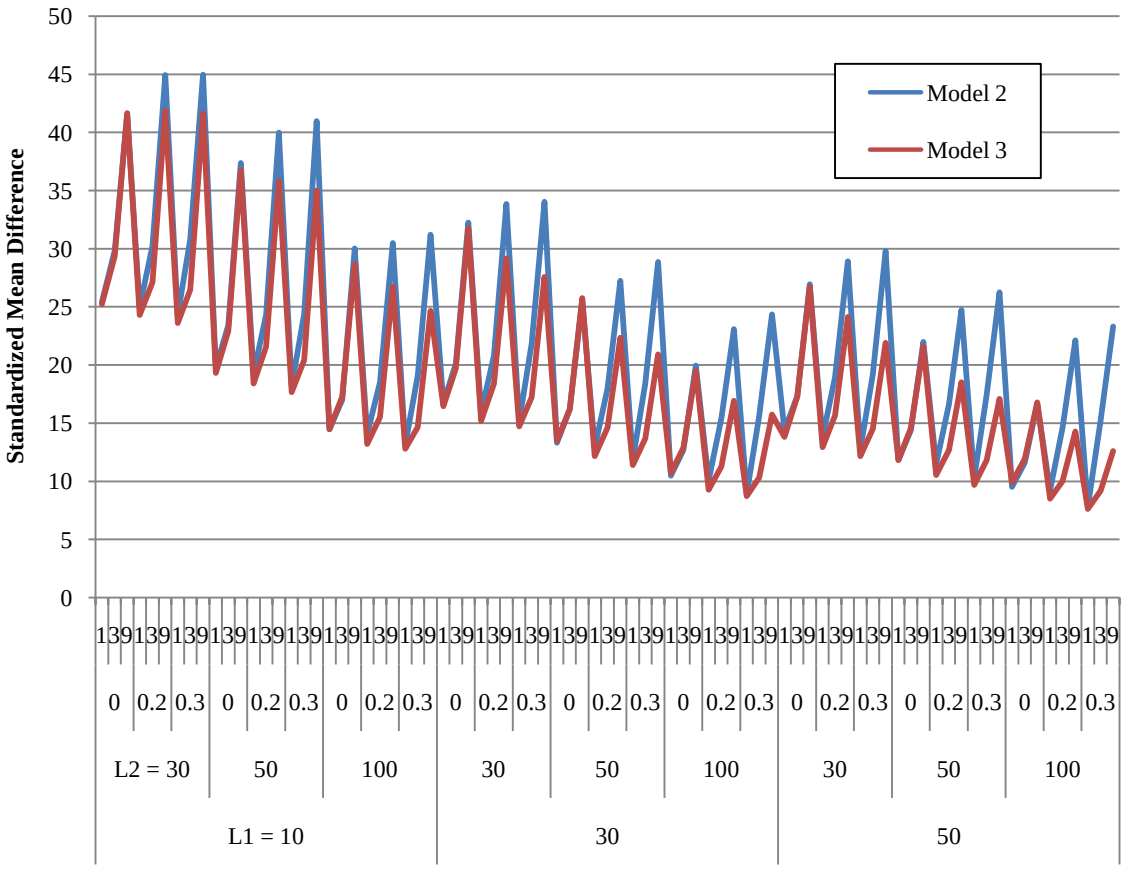
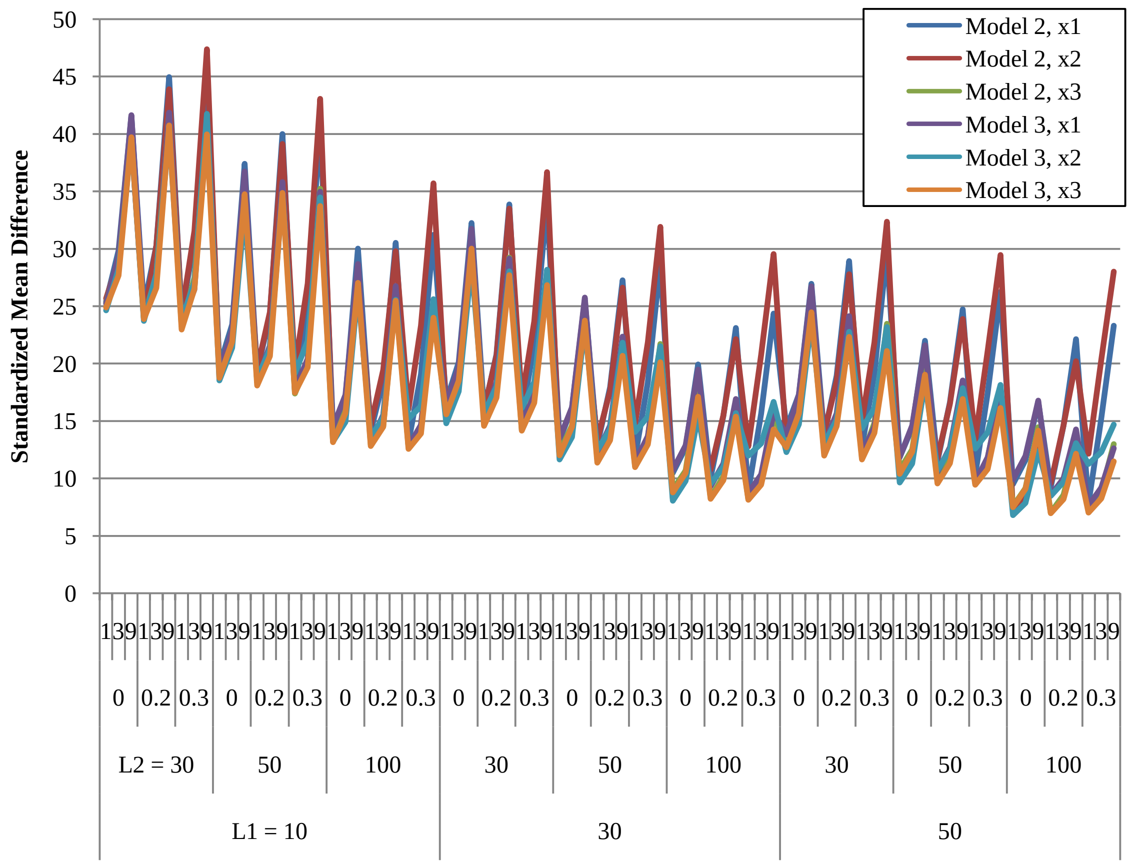


Figure 85: SMD for X_1, X_2, and X_3 across Simulation Conditions for Quintile Stratification with Propensity Scores Estimated using Model 2 and Model 3



SMD versus Mean γ_{10}

The patterns apparent in the SMD and the mean γ_{10} for quintile stratification show both similarities and differences. Both show similar relationships with treatment-control group ratio, indicating poorer balance when the ratio is 1:9 than when it is 1:3 or 1:1. The patterns were dissimilar when considering sample-size: SMD showed better balance with larger sample sizes whereas mean γ_{10} indicated poorer balance with larger sample sizes. The poorer balance for the mean γ_{10} was primarily limited to the 1:9 ratio condition. When using Model 2, SMD showed poorer balance with stronger cross-level interactions whereas mean γ_{10} indicated better balance. When W was not included in the propensity

score estimation, however, both SMD and mean γ_{10} showed better balance with stronger cross-level interactions. As discussed previously, this pattern is likely related to the fact that SMD does not take into account the clustering of the data when estimating variance but assumes that the variance is homogenous across clusters. A potential explanation for this finding is that, when W is not included in the propensity score estimation, larger cross-level interactions would result in greater variance in the covariates between treatment and control groups and smaller subsequent SMDs. When W is included in the propensity score estimation, the cross-level interactions are better addressed through the stratification method. The variance in the covariates between groups is also better addressed and would not influence the values of the SMD as strongly.

Figure 86: Overlay of SMD and Mean γ_{10} for Quintile Stratification with Propensity Scores Estimated using Model 2

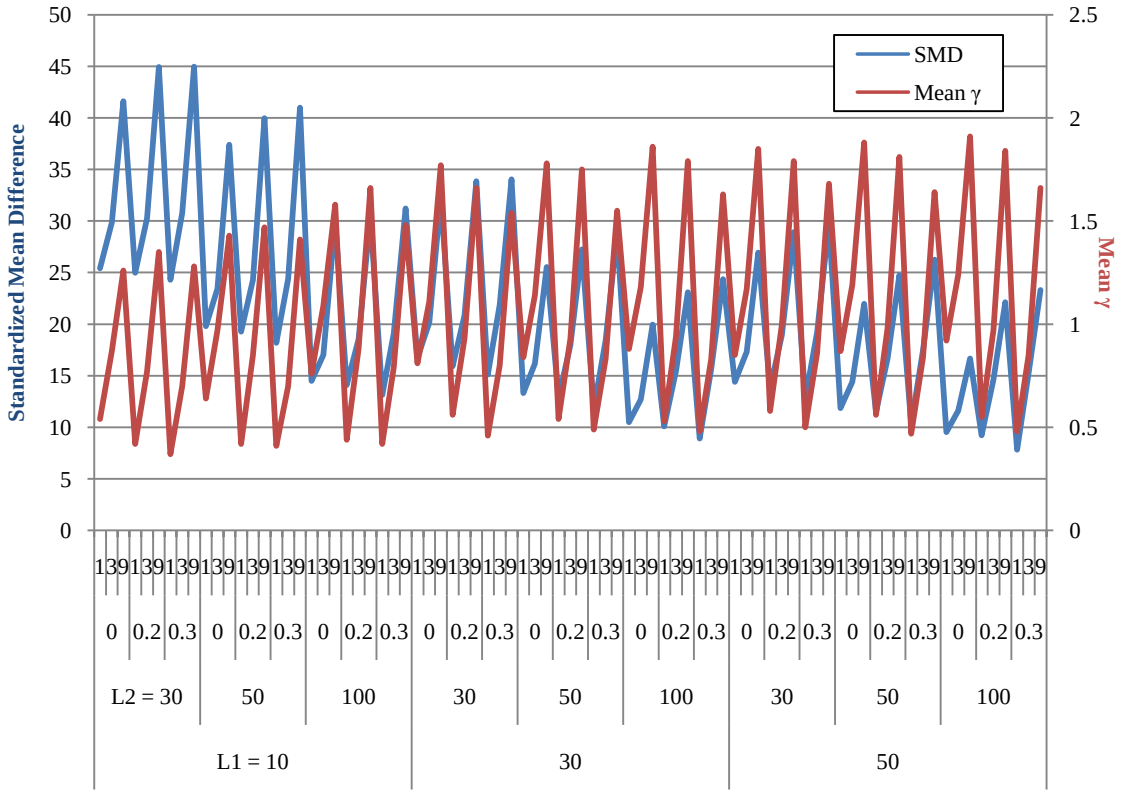
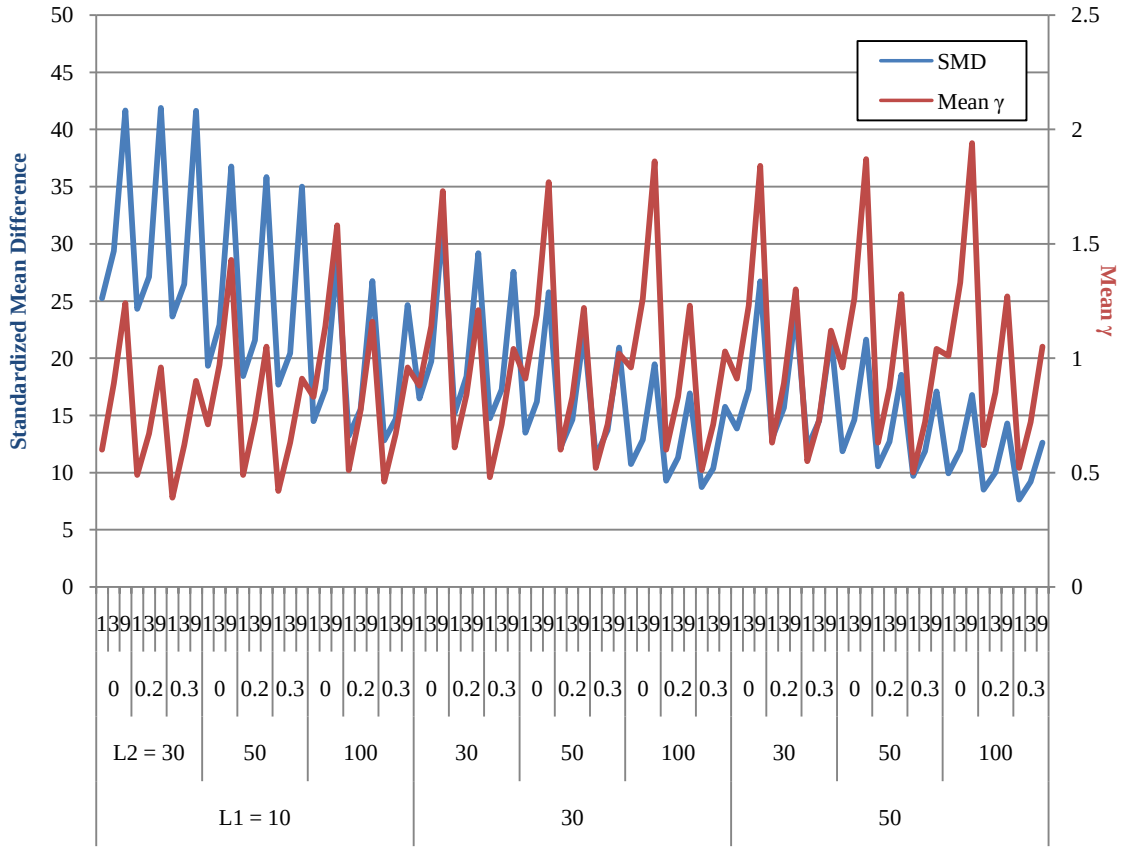


Figure 87: Overlay of SMD and Mean γ_{10} for Quintile Stratification with Propensity Scores Estimated using Model 3



Comparison of the Performance of the Adjustment Methods using Propensity Scores Estimated through Logistic Regression

Retaining Treatment Group Members. The adjustment method that resulted in the greatest percentage of retained treatment group members was quintile stratification, followed by between-cluster matching, and within-cluster matching. This pattern was consistent across both Model 2 and Model 3. The propensity scores estimated using Model 3 retained slightly more treatment group members than did the propensity scores estimated using Model 2 (average of .3% more for quintile stratification, 2.2% more for between-cluster matching, and .6% more for within-cluster matching).

Sample size retention rates show a positive relationship with the level-1 sample size. This relationship is strongest when using within-cluster matching. A positive relationship is also apparent between sample size retention rates and the level-2 sample size; however, this relationship is small. These relationships are consistent across propensity score Models 2 and 3. The average retention rates per sample-size condition are presented in Table 48 below.

Table 48:

Percent of the Initial Treatment Group Retained per Propensity Score Adjustment Method for Propensity Scores Estimated using Model 2 and Model 3

Level- 2 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
Model 2										
30	22.2	39.7	47.2	75.3	78.4	79.0	94.4	97.7	98.5	70.3
50	22.2	39.9	47.5	78.1	79.9	80.4	96.7	98.7	99.2	71.4
100	22.4	39.9	47.6	80.2	81.1	81.3	98.3	99.3	99.6	72.2
Mean	22.3	39.8	47.4	77.9	79.8	80.3	96.4	98.6	99.1	71.3
Model 3										
30	22.6	40.3	47.8	77.8	80.6	81.2	95.1	98.1	98.8	71.4
50	22.7	40.5	48.1	80.5	82.0	82.5	97.2	98.9	99.3	72.4
100	22.9	40.5	48.2	82.4	83.1	83.3	98.6	99.5	99.7	73.1
Mean	22.7	40.4	48.0	80.2	81.9	82.3	97.0	98.9	99.3	72.3

The relationship between the percent of the treatment group retained and the cross-level interaction is small and positive: As the cross-level interaction increases, the percent retained increases. Both Models 2 and 3 show almost identical patterns in this relationship, although, as the cross-level interactions increases from 0 to 0.2 and 0.3, Model 3 retained slightly higher percentages of treatment group members in the 1:9 ratio condition than in the 1:3 and 1:1 ratio conditions.

A strong relationship is apparent between the treatment-group ratio and the percent treatment group retained when using matching methods: As the control group members increase in proportion to the treatment group members, the percent retained increases. A different pattern is apparent when applying a quintile stratification adjustment method, in which the moderate treatment-control group ratio shows the highest percent retained, although the difference between the percent retained per condition are very small across cells. These relationships are illustrated in Figure 88 below.

Figure 88: Percent treatment group retained per cross-level interaction and treatment-control group ratio per propensity score adjustment method and estimation model

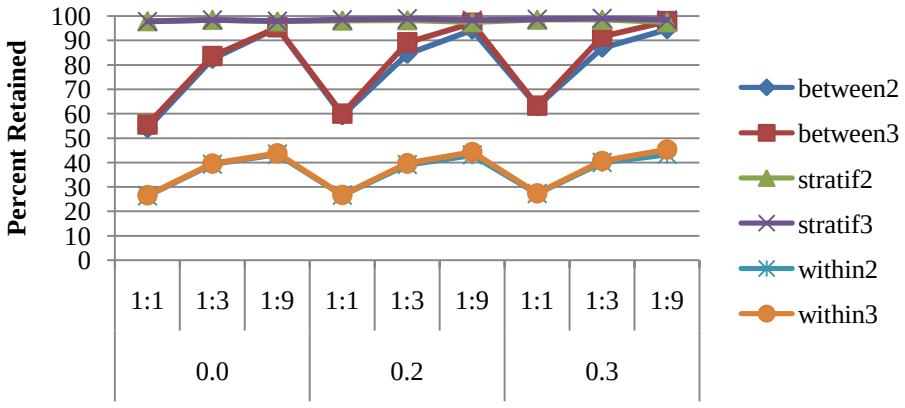
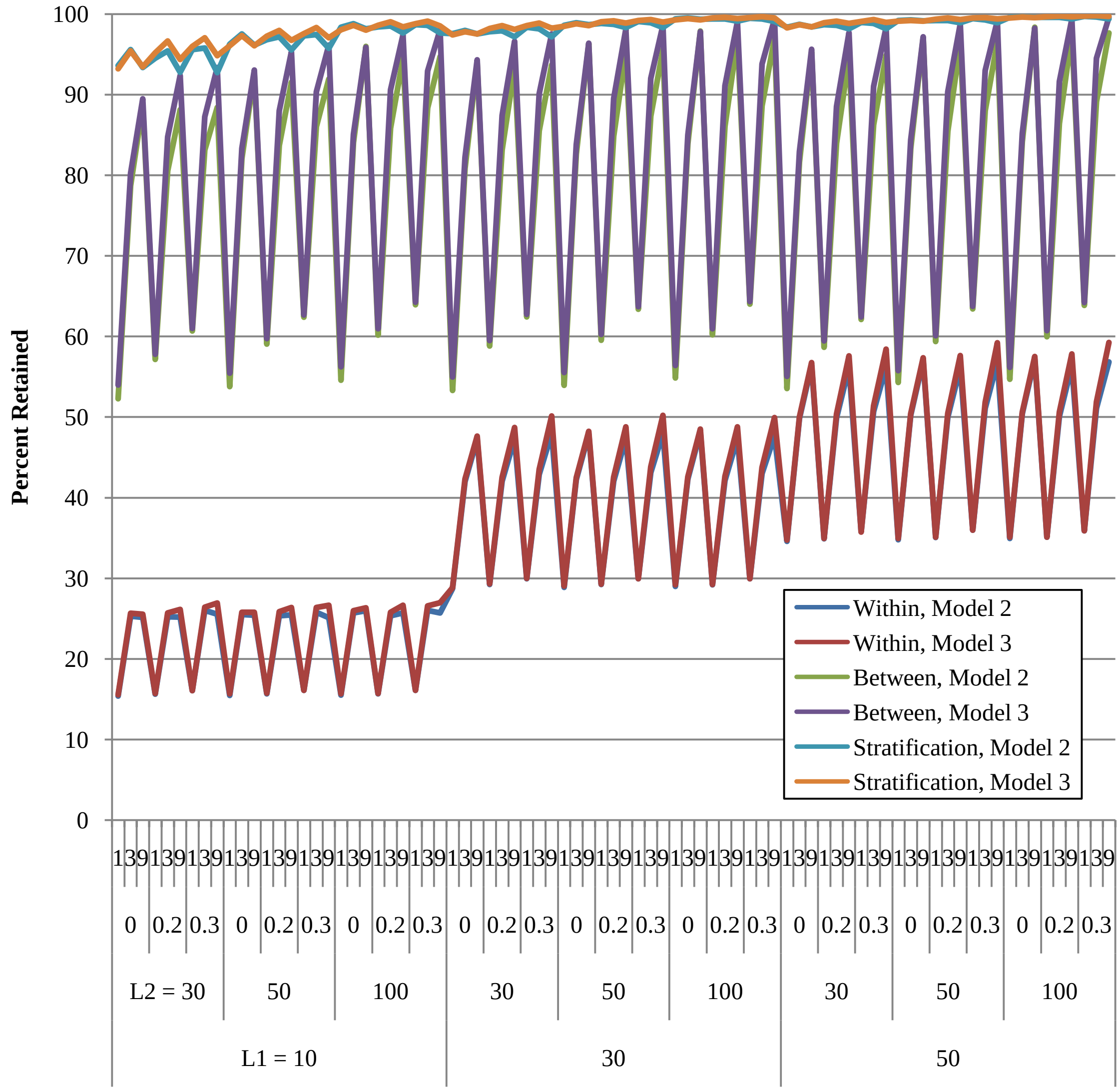


Figure 89: Percent of Sample Retained per Simulation Condition and Adjustment Method for Propensity Scores Estimated using Method 2 and Method 3



Cluster-Level Sample Retention. When considering the percent of clusters that were retained per propensity score estimation model and adjustment method, within-cluster matching resulted in the smallest percent of clusters retained. Between-cluster matching and quintile stratification showed high percentages of clusters retained, with all cells retaining over 90%. Across conditions, the level-2 sample size had little bearing

upon the percent of the clusters that were retained. Within-cluster matching was the only propensity score adjustment method that showed a positive relationship with the level-1 sample size. For both between-cluster matching and quintile stratification, the percent of clusters retained typically decreased as the level-1 sample size increased, although this decrease was small (see Table 49).

Table 49:

Percent of the Clusters Retained per Propensity Score Adjustment Method and Sample Size Conditions

Level-2 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification			Mean
	Level-1 Sample Size			Level-1 Sample Size			Level-1 Sample Size			
	10	30	50	10	30	50	10	30	50	
Model 2										
30	45.9	87.9	93.3	96.5	97.1	96.1	98.3	97.1	96.1	89.8
50	46.1	88.0	93.3	97.1	97.1	96.1	98.3	97.1	96.1	89.9
100	46.2	88.0	93.4	97.6	97.1	96.2	98.3	97.1	96.2	90.0
Mean	46.1	88.0	93.4	97.1	97.1	96.1	98.3	97.1	96.1	89.9
Model 3										
30	46.5	88.3	93.5	97.1	97.1	96.1	98.3	97.1	96.1	90.0
50	46.7	88.4	93.5	97.6	97.1	96.1	98.3	97.1	96.1	90.1
100	46.8	88.5	93.7	97.9	97.1	96.2	98.3	97.1	96.2	90.2
Mean	46.7	88.4	93.6	97.5	97.1	96.1	98.3	97.1	96.1	90.1

Across propensity-score adjustment methods, a small positive relationship is apparent between the percent clusters retained and the cross-level interaction. The treatment-control group ratio has differential relationships with the percent clusters retained across the adjustment methods. For within-cluster matching, the percent of the clusters retained is smallest in the 1:9 ratio condition as compared to the 1:3 and 1:1 ratio conditions, which are very similar. These relationships are illustrated in Figure 90 below.

Figure 90: Percent Clusters Retained per Cross-Level Interaction and Treatment-Control Group Ratio for Adjustment Methods and Estimation Models 2 and 3

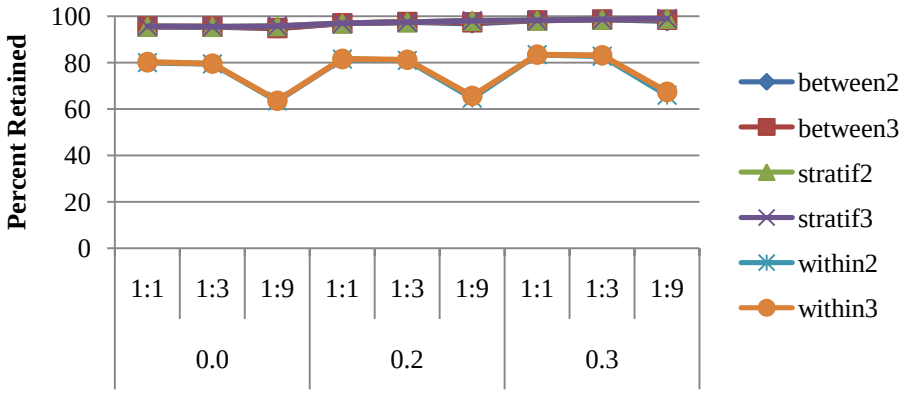
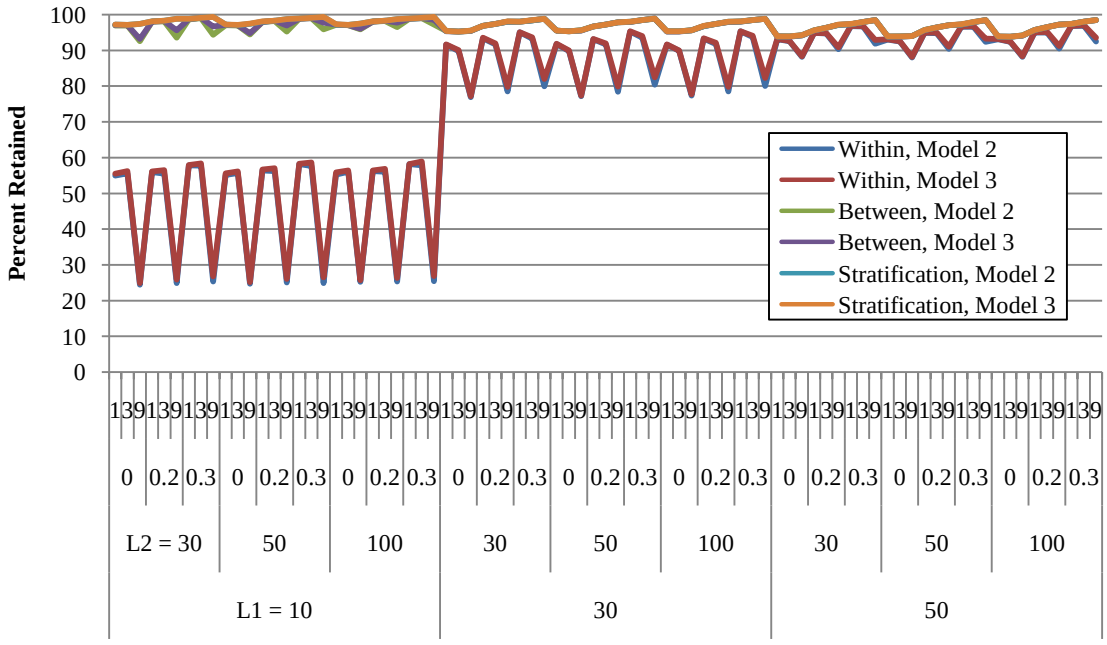


Figure 91: Percent of the Initial Number of Clusters Retained across Simulation Conditions per Adjustment Method and Propensity Score Estimation Models 2 and 3



Percent Significant γ_{10} across Propensity Score Adjustment Methods

The average percent of significant γ_{10} was small across all simulation conditions for propensity scores estimated using Models 2 and 3 and applied using the three adjustment methods for all three covariates (see Table 50). When considering the balance achieved across covariates, average percent of significant γ_{10} across all conditions for X_2 and X_3 remained below 1%.

When examining overall success in balancing covariates, between-cluster matching showed the smallest overall average percent of significant γ_{10}; quintile stratification showed the highest average percent of significant γ_{10}. Between-cluster matching showed smaller average percentage of significant γ_{10} when propensity scores were estimated using Model 2 versus Model 3. When within-cluster matching was used,

little difference was apparent between propensity score estimation models. The largest average percent of significant γ_{10} was apparent when quintile stratification was used. Quintile stratification also showed the greatest range in the percent of significant γ_{10} across cells compared to other adjustment methods, with a high of 61% when the propensity score was estimated using Model 2 (versus a high of 7% for within-cluster matching and 10% for between-cluster matching) and 53% when the propensity score was estimated using Model 3 (versus a high of 5% for within-cluster matching and 20% for between-cluster matching).

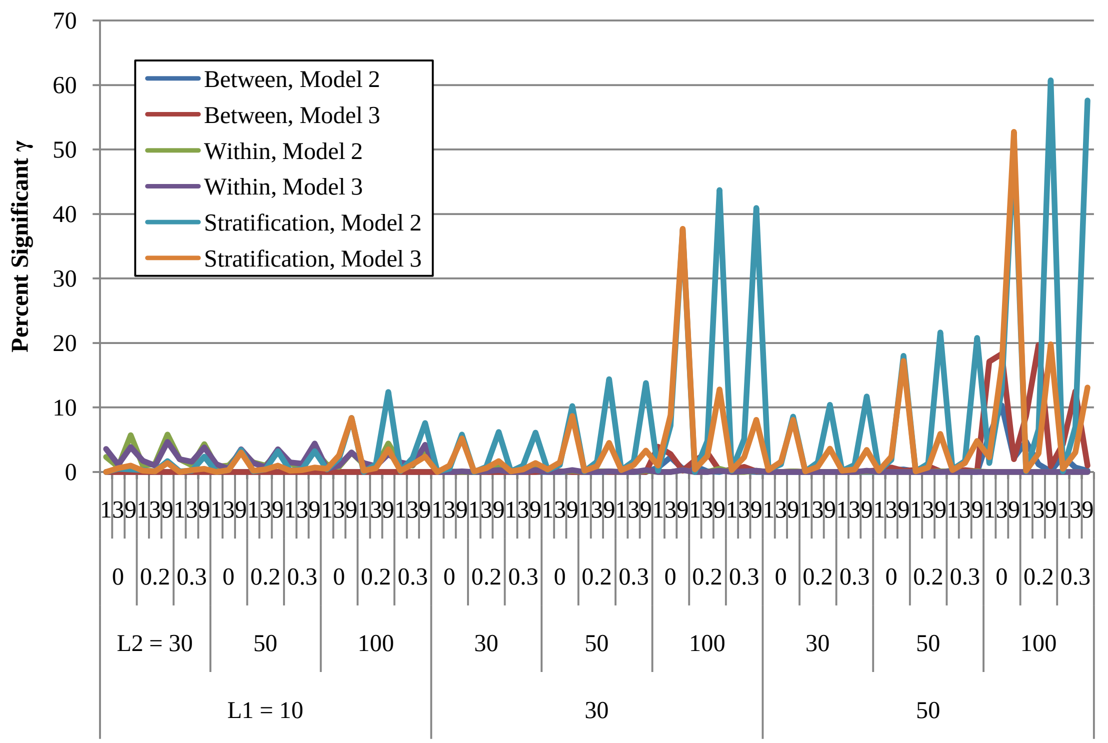
When total sample size was less than 1000, all methods maintained percent significant γ_{10} below 10% regardless of other conditions. As sample-size surpassed 1000, however, the quintile stratification methods show percent of significant γ_{10} above 10% when the treatment-control group ratios were 1:9. Between-cluster matching did not show increases in percent of significant γ_{10} until the sample size reached 3000 (level-1 n = 30 and level-2 n = 100), although these increases were not apparent in the 1:9 treatment-control ratio condition. The values for the percent of significant γ_{10} for X_1 across conditions are presented in Figure 92 below.

Table 50:

The Average Percent of Significant γ_{10} across Xs, Propensity Score Adjustment Models, and Propensity Score Estimation Models 2 and 3

Predictor	Within-Cluster Matching		Between-Cluster Matching		Quintile Stratification	
	Model 2	Model 3	Model 2	Model 3	Model 2	Model 3
X_1	0.75	0.74	0.41	1.25	6.83	3.67
X_2	0.91	0.82	0.03	0.06	0.12	0.02
X_3	0.79	0.92	0.01	0.01	0.00	0.00
Mean	0.82	0.83	0.15	0.44	2.32	1.23

Figure 92: Percent Significant γ_{10} across Simulation Conditions, Adjustment Methods, and Propensity Score Estimation Models 2 and 3



Percent Significant τ_{11} across Propensity Score Adjustment Methods

When averaged across conditions, the variance in the balance achieved across clusters, as indicated by significant values of τ_{11}, is above 50% for all adjustment methods and propensity score estimation models (see Table 51). For between-cluster matching and quintile stratification, the percent significant τ_{11} is consistently above 90%. Results related to within-cluster matching are smaller than other adjustment methods across simulation conditions. The percent significant τ_{11} also shows a negative relationship with the strength of the relationship between X and Z: The percent significant τ_{11} is higher for X_3 as compared to X_2 and X_1. The regression-based propensity score estimation models show almost identical results in the average percent significant τ_{11} per adjustment method.

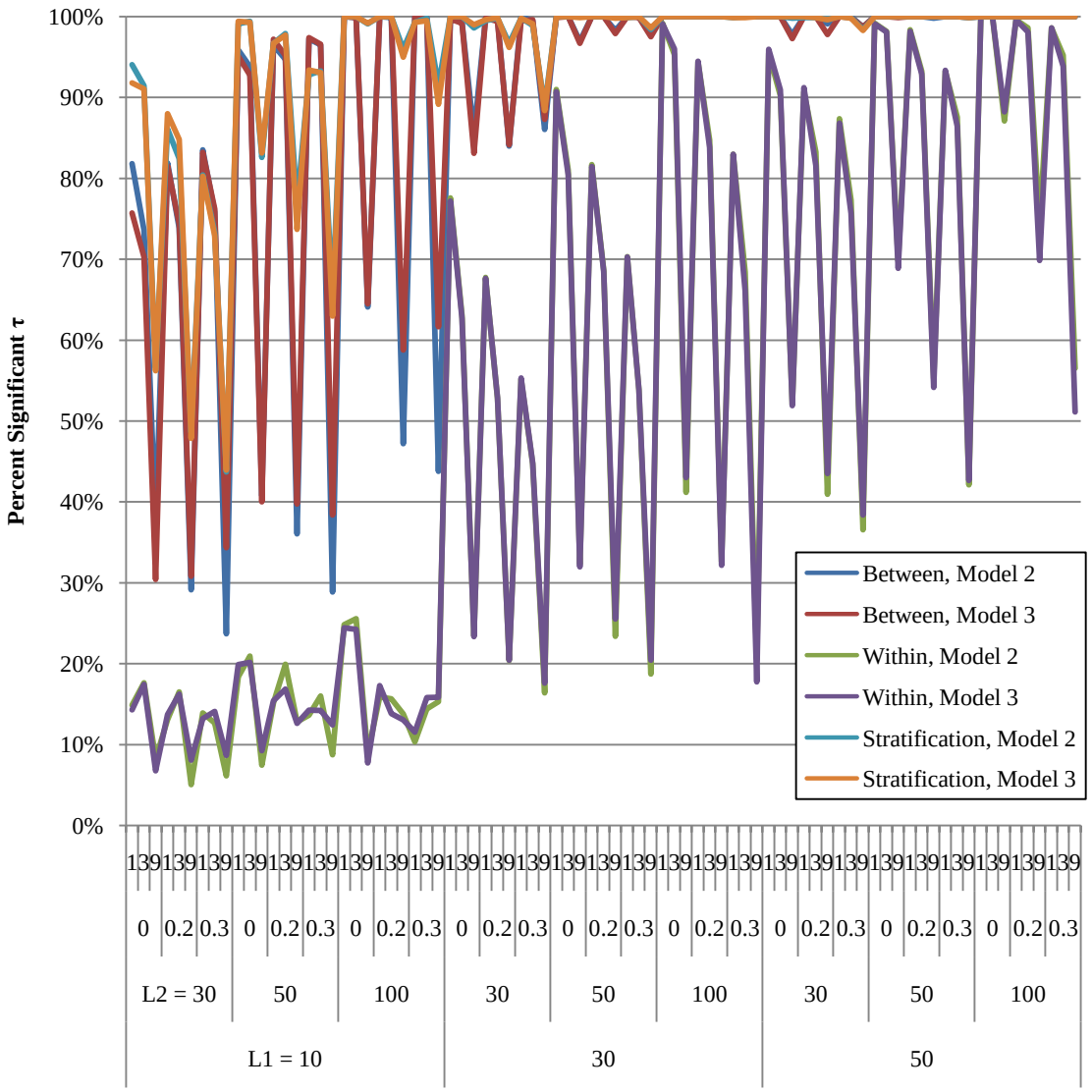
The relationships of the cross-level interaction and of the treatment-control ratio with percent significant τ_{11} are apparent in Figure 93 below. Across sample-size conditions, adjustment methods, and propensity score models, these relationships are both strong and negative: As cross-level interactions increase in strength, the percent significant τ_{11} decreases; and as the control group members become greater in proportion to the treatment group members, the percent significant τ_{11} decreases.

Table 51:

The Average Percent of Significant τ_{11} across Xs, Propensity Score Adjustment Method, and Logistic Regression Estimation Models

Predictor	Within-Cluster Matching		Between-Cluster Matching		Quintile Stratification	
	Model 2	Model 3	Model 2	Model 3	Model 2	Model 3
X_1	51	51	91	91	95	95
X_2	74	74	93	92	96	96
X_3	82	82	94	94	97	98
Mean	69	69	92	92	96	96

Figure 93: Percent Significant τ_{11} across Simulation Conditions per Adjustment Methods and Propensity Score Estimation Model for X_1

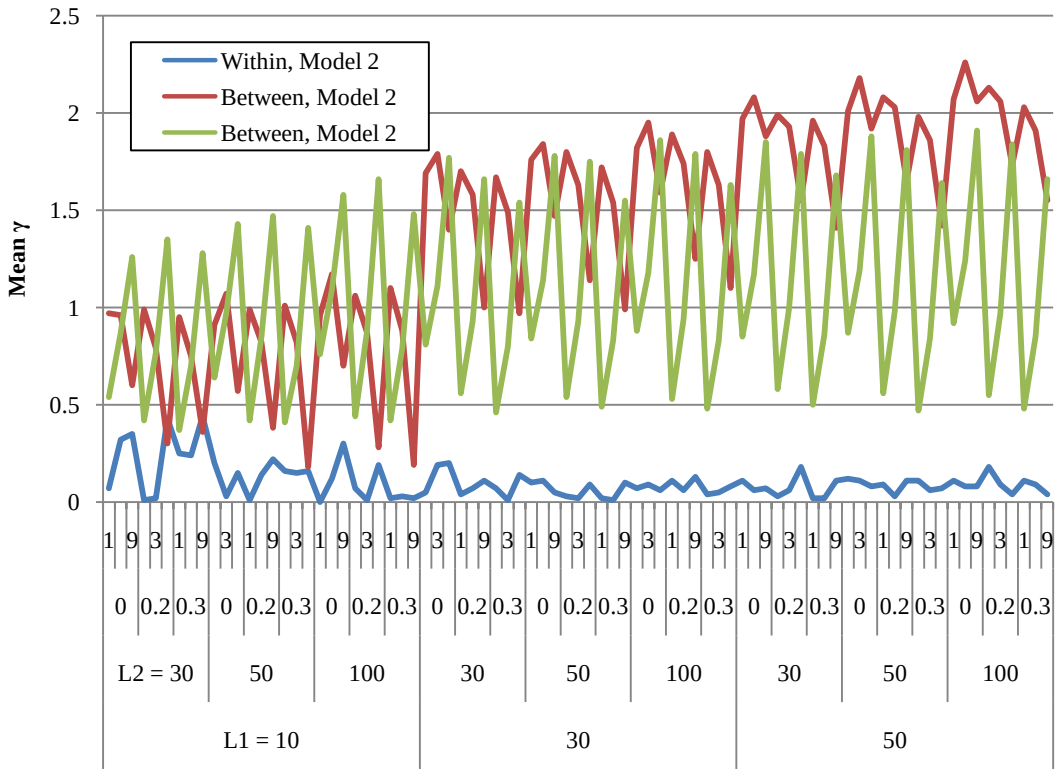


Comparison of Mean γ_{10} and Mean τ_{11} per Adjustment Method

The adjustment method that resulted in the smallest mean values of γ_{10} across conditions was within-cluster matching. Between-cluster matching and quintile stratification showed similar results, although quintile stratification showed a slight advantage when sample size at level-1 was moderate or large. Considering the treatment-

control ratio, between-cluster matching and quintile stratification clearly showed opposite patterns of balance: Between-cluster matching resulted in better balance in the 1:9 ratio condition, whereas quintile stratification showed poorer balance in the 1:9 ratio condition. Although less apparent in Figure 94, within-cluster matching also resulted in poorer matches in the 1:9 condition. The benefit that between-cluster matching evidences in the 1:9 condition is likely due to the larger control reservoir from which to find close matches which is not available to within-cluster matching methods.

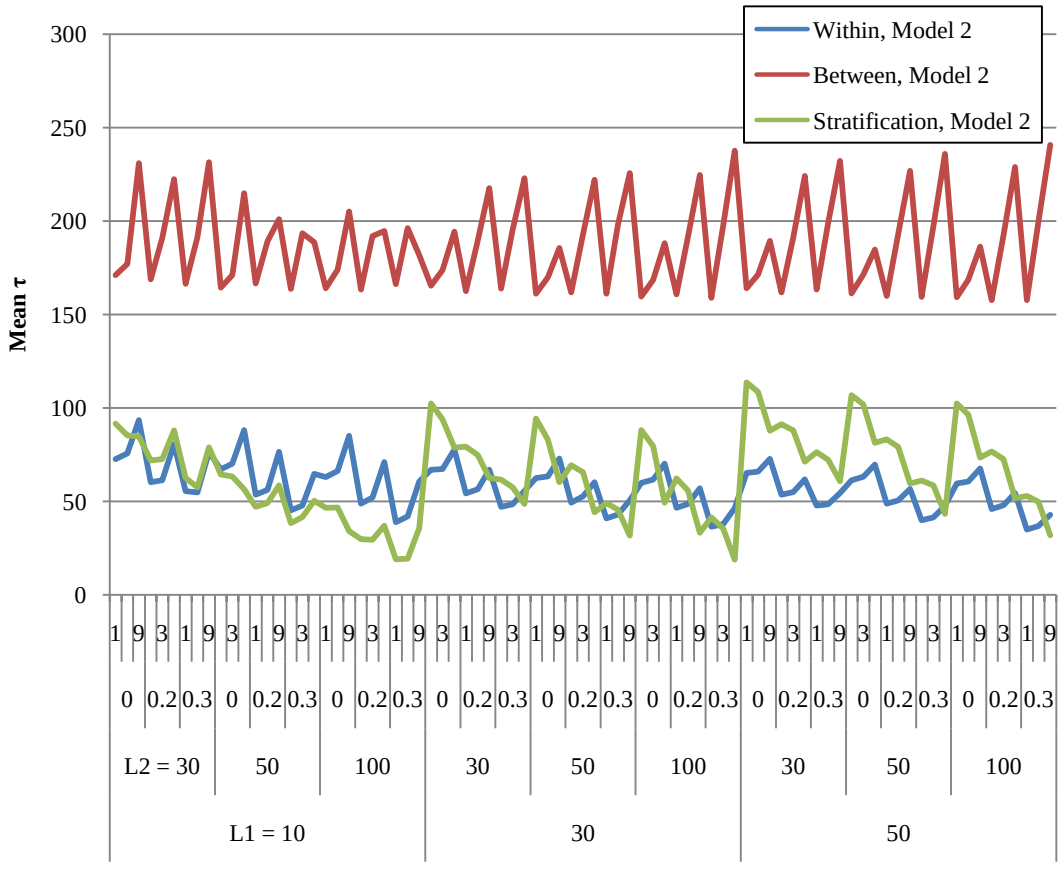
Figure 94: Mean γ_{10} across Adjustment Methods for Propensity Scores Estimated using Model 2



The variance in the balance across clusters that results from within-cluster matching and quintile stratification are similar across simulation conditions, although

quintile stratification evidences relatively larger fluctuations in values of τ_{11} across conditions. Between-cluster matching, however, shows the highest values of τ_{11} than do other adjustment methods. As noted previously, matching between clusters may lead to clusters with only treatment group members or only control group members resulting in dissimilar individuals within each cluster. When quintile stratification is applied, both treatment and control group members are far more likely to be present in each cluster than when between-cluster matching is applied.

Figure 95: Comparison of Mean τ_{11} across Adjustment Methods for Propensity Scores Estimated using Model 2



Standardized Mean Difference across Propensity Score Adjustment Methods

Between-cluster matching resulted in smaller average SMD than did other adjustment methods when applying logistic regression-based propensity score estimation models. Within-cluster matching is consistently the moderate performer across conditions while quintile stratification results in the largest average SMDs. Once overall sample size reaches 1500, both within-cluster matching and between-cluster matching result in SMDs below 10%, regardless of other simulation conditions. When a cluster-level predictor is included in the estimation model, the average SMD is slightly higher when between-cluster matching and quintile stratification is used compared to results from a propensity score estimated without a cluster-level predictor. No difference in average SMD is apparent between propensity score estimation models when within-cluster matching is used. The average SMD is consistently smaller for X_1 than for X_2 and X_3 when matching methods are used. When quintile stratification was applied, however, results were inconsistent across Xs.

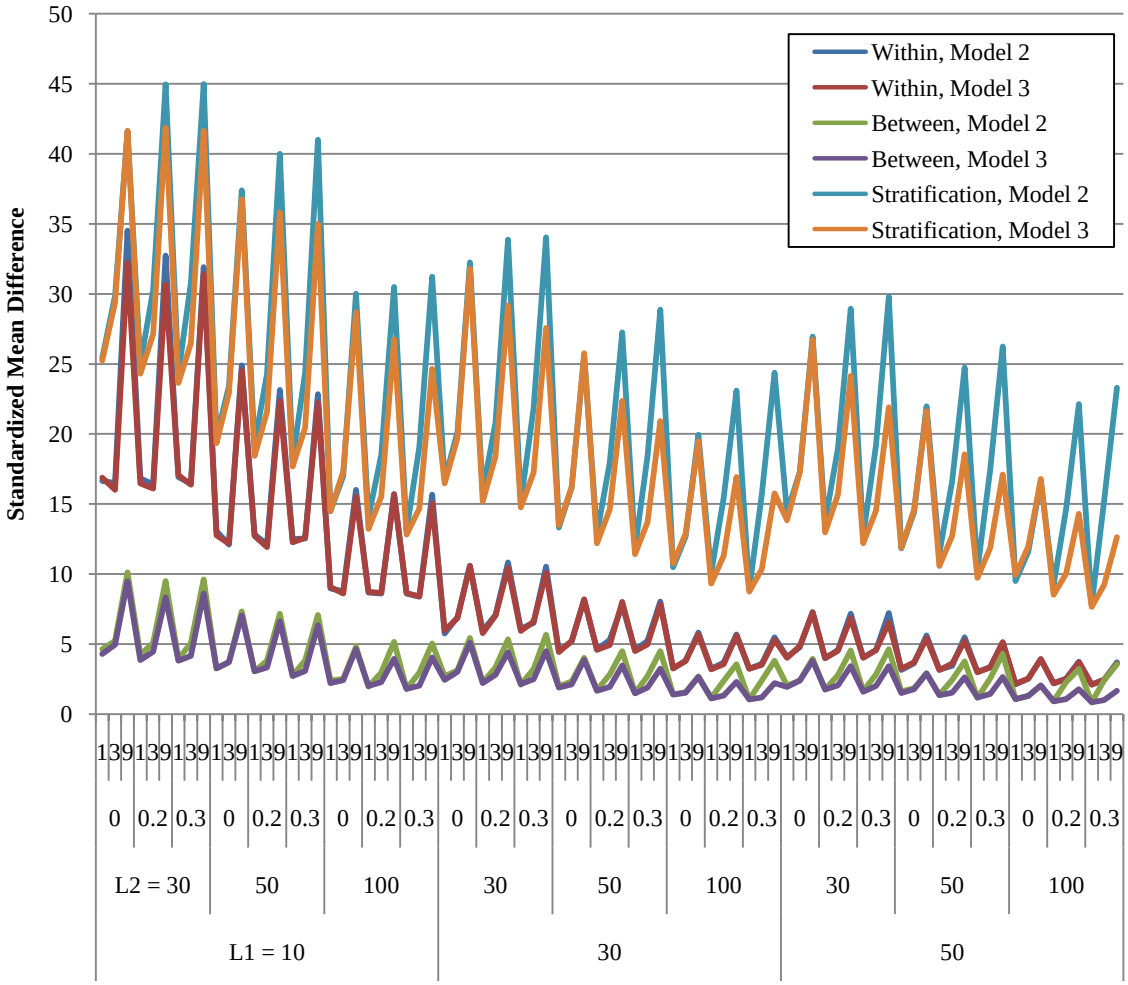
The range in the SMD across cells per propensity score estimation models was very similar per propensity score estimation method, and therefore will be discussed together. The smallest range was apparent when between-cluster matching was applied (0.83 - 10.1), followed by within-cluster matching (2.1 - 31.5), and quintile stratification (7.6 - 45.0). The SMDs across simulation conditions for regression-based propensity score estimation models are presented in Figure 96 below.

Table 52:

The Average SMD across Xs, Propensity Score Adjustment Methods, and Logistic Regression Estimation Models

	Within-Cluster Matching		Between-Cluster Matching		Quintile Stratification	
Predictor	Model 2	Model 3	Model 2	Model 3	Model 2	Model 3
X_1	8.8	8.7	3.4	2.9	21.3	18.8
X_2	16.3	16.3	6.1	5.9	21.5	13.4
X_3	14.3	14.2	5.2	5.0	18.2	17.6
Mean	13.1	13.1	4.9	4.6	20.3	16.6

Figure 96: SMD for X_1 Propensity Score Adjustment Method and Estimation Models 2 and 3



Comparison of All Propensity Score Models and Adjustment Methods

Balance across the Sample. The performance of each propensity score and adjustment method is discussed in the following section. Overall, the combination of propensity score estimation model and adjustment method that resulted in the smallest mean value of γ_{10} was within-cluster matching using the most basic of the propensity scores: Model 3, which included no cluster-level predictor. This method, however, did not result in consistent balance across clusters, as indicated by the mean values of τ_{11}.

Under most simulation conditions, the propensity score estimated using a multilevel model showed larger mean values of γ_{10} than other propensity score estimation models within adjustment method. The difference in the performance of this propensity score compared to other scores, however, was diminished (in the case of within-cluster matching) or removed (in the case of between-cluster matching and quintile stratification) once sample size at level-1 reached 50.

When considering the balance achieved across covariates, propensity scores estimated using Model 1 show better balance with smaller covariate correlations with treatment assignment (mean value of γ_{10} for $X_3 < X_2 < X_1$). Under most conditions, propensity scores estimated using Model 3 show better balance with larger covariate correlations with treatment assignment (mean value of γ_{10} for $X_3 > X_2 > X_1$), although this pattern is not consistent when between-cluster matching is used. Propensity scores estimated using Model 2 show inconsistent balance patterns across covariates across adjustment methods.

Table 53:

Average Mean γ_{10} for X_1, X_2, and X_3 across Propensity Scores and Adjustment Methods

Level-1 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
				X_1					
10	4.44	0.15	0.12	3.53	0.76	0.91	3.08	0.92	0.82
30	2.14	0.08	0.07	2.86	1.55	1.74	2.19	1.10	0.98
50	1.21	0.08	0.07	2.03	1.91	2.11	1.17	1.15	1.02
Mean	2.60	0.10	0.09	2.81	1.41	1.59	2.15	1.06	0.94
				X_2					
10	3.90	0.35	0.54	3.16	1.00	1.15	2.65	1.15	1.09
30	1.74	0.44	0.59	2.44	1.85	2.03	1.87	1.42	1.36
50	1.00	0.55	0.71	1.71	2.21	2.46	0.99	1.52	1.47
Mean	2.21	0.45	0.61	2.44	1.69	1.88	1.84	1.36	1.31
				X_3					
10	2.92	0.98	1.17	1.89	0.83	0.84	1.75	0.86	1.07
30	1.45	1.61	2.17	1.78	1.18	1.23	1.46	1.19	1.46
50	0.85	2.05	2.62	1.47	1.30	1.42	1.06	1.32	1.61
Mean	1.74	1.55	1.99	1.71	1.10	1.16	1.42	1.12	1.38

Discussion of Indicators of Overall Balance. Under most conditions, the SMD provided indications of balance that were similar to those apparent from the mean values of γ_{10}. The poorer performance in balancing X_1 across the sample as a whole that was

evident from application of the propensity scores estimated using Model 1, compared to other propensity scores models, was much more evident in the SMD. Additionally, the values for the SMD resultant from quintile stratification showed greater imbalances than were apparent in the mean value of γ_{10}. As noted previously, the difference in the results from the mean value of γ_{10} and the SMD are likely resultant from the ability of each indicator to take into account the effects of clustering. The application of SMD to determine balance in data with a multilevel structure might lead to overestimation of the imbalances in covariates between treatment and control groups. The performance of the percent significant γ_{10} showed great sensitivity to sample size and is, therefore, a misleading indicator when estimating balance that is achieved across simulation replications when sample sizes are large. Mean γ_{10} appears to be a reliable indicator of balance when sufficient sample size exists in each cluster to estimate cluster-level effects.

Table 54:

SMD across Propensity Scores and Adjustment Methods

Level-1 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
10	55.2	16.5	16.1	37.4	4.8	4.3	50.3	27.1	25.1
30	24.5	6.0	5.9	28.7	2.9	2.4	39.5	19.4	16.9
50	16.6	4.0	4.0	23.5	2.4	1.9	34.5	17.3	14.4
Mean	32.1	8.8	8.7	29.8	3.4	2.9	41.5	21.3	18.8

The specific values for these measures of balance at each simulation condition are also useful to consider. In the following three figures, the values for mean γ_{10} resulting from each estimation model are compared. For within-cluster matching, the results show that propensity scores estimated using Model 1 are never as successful in balancing covariates across the sample as those estimated from Models 2 and 3. This finding is consistent for within-cluster matching, regardless of the number of individuals per cluster or the number of clusters. When between-cluster matching and quintile stratification are applied, however, the balance attained through the application of propensity scores from Model 1 is equal to or better than that attained by the application of scores from Models 2 and 3.

Results from Model 1 are also more influenced by the cross-level interaction than those from Models 2 and 3, while results from Models 2 and 3 are more influenced by the treatment-control group ratio than those from Model 1. The relationship of balance and treatment-control group ratio is particularly apparent in quintile stratification (Figure 99), and, to a lesser degree, in between-cluster matching (Figure 98). The following figures also indicate that these benefits associated with Model 1 are limited to the largest sample size conditions where these propensity scores can be efficiently estimated.

Figure 97: Mean γ_{10} of X_1 across Propensity Score Estimation Methods when using Within-Cluster Matching

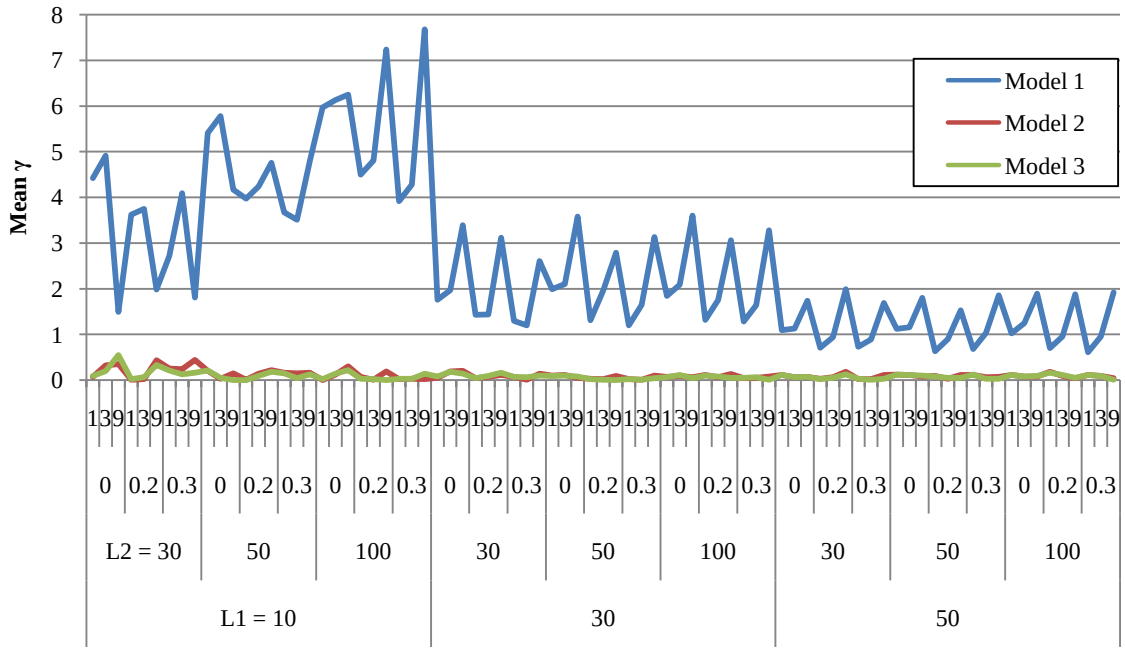


Figure 98: Mean γ_{10} of X_1 across Propensity Score Estimation Methods when using Between-Cluster Matching

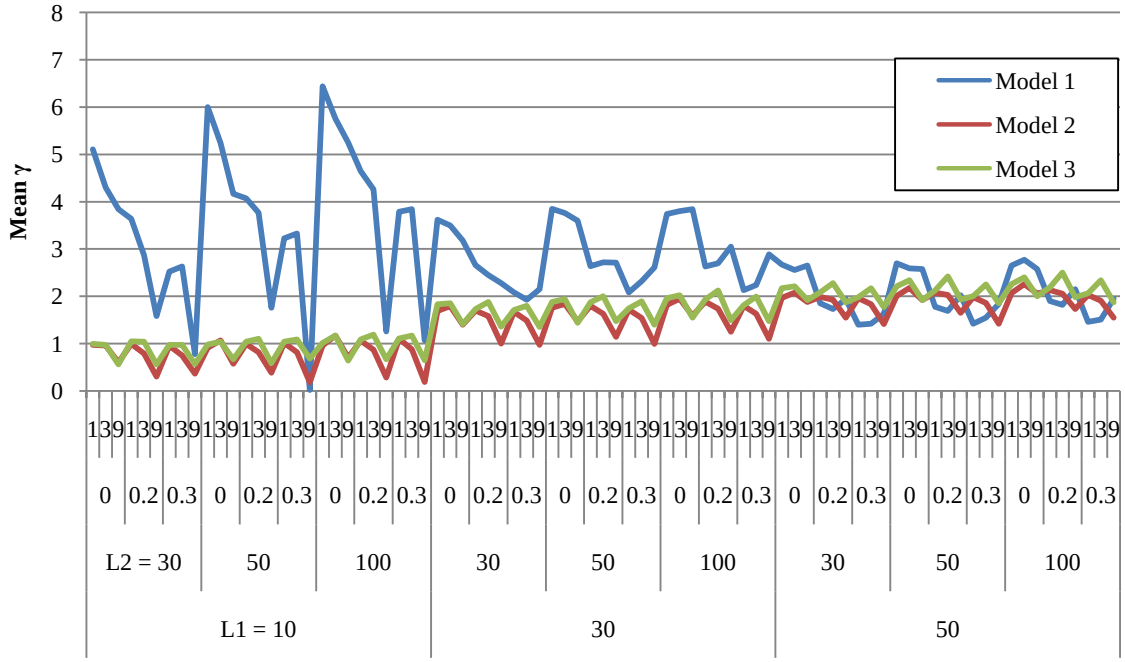
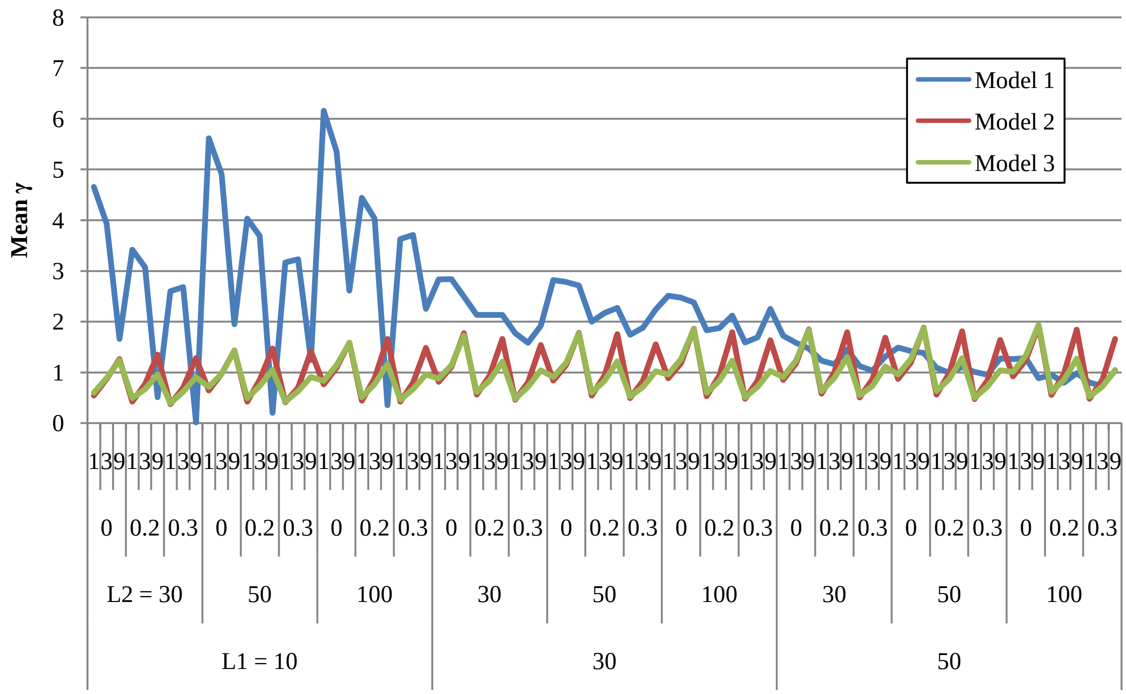


Figure 99: Mean γ_{10} of X_1 across Propensity Score Estimation Methods when using Quintile Stratification



Balance across Clusters. The combination of adjustment method and propensity score estimation model that resulted in the smallest variance in balance across clusters (τ_{11}) was within-cluster matching when using the propensity score estimated using a multilevel model. Within each adjustment method, the propensity score estimated using Model 1 resulted in less cross-cluster variance than propensity scores that were estimated using logistic regression. The one exception to this pattern, when other propensity scores outperformed the propensity score from Model 1, was when the scores were applied using quintile stratification and sample size at level-1 was at its smallest.

The variance in the balance across clusters increases greatly with increased covariate correlation with treatment assignment. Variance is consistently greater across covariates when using propensity scores estimated from Model 2 or Model 3 compared to

Model 1. When using a propensity score estimated using Model 2 or Model 3, the mean values for τ_{11} are smallest for X_1 and largest for X_3. This pattern is not consistent for a propensity score estimated using Model 1. When quintile stratification is used with Model 1, X_2 shows slightly less variation in balance than does X_1. When applied to within-cluster matching and between-cluster matching, however, the variance is largest in X_2.

Table 55:

Average Mean τ_{11} for X_1, X_2, and X_3 across Propensity Scores and Adjustment Methods

Level-1 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
				X_1					
10	40.0	64.0	63.1	45.7	186.7	180.0	67.3	54.0	52.9
30	5.2	55.6	55.3	15.6	187.1	174.4	44.1	61.7	60.9
50	7.8	53.9	53.8	16.4	188.1	174.9	33.3	76.0	75.0
Mean	17.67	57.83	57.4	25.9	187.3	176.43	48.23	63.9	62.93
				X_2					
10	177.8	461.9	458.5	109.1	592.2	786.4	84.4	326.4	327.9
30	125.0	537.6	541.7	35.1	563.3	532.9	8.4	308.1	310.0
50	93.9	558.2	560.9	40.2	562.2	534.3	16.9	301.2	303.0
Mean	132.2	519.2	520.4	61.5	572.6	617.9	36.5	311.9	313.6
				X_3					
10	146.9	1030.6	959.3	124.2	766.3	773.9	363.2	1201.9	1190.0
30	50.9	968.8	920.0	11.2	749.0	758.3	191.2	1412.1	1407.9
50	60.2	956.7	916.0	24.5	738.1	747.7	146.9	1416.6	1414.8
Mean	86.0	985.4	931.8	53.3	751.1	760.0	233.8	1343.5	1337.6

Figure 100: Mean τ_{11} of X_1 across Propensity Score Estimation Methods when using Within-Cluster Matching

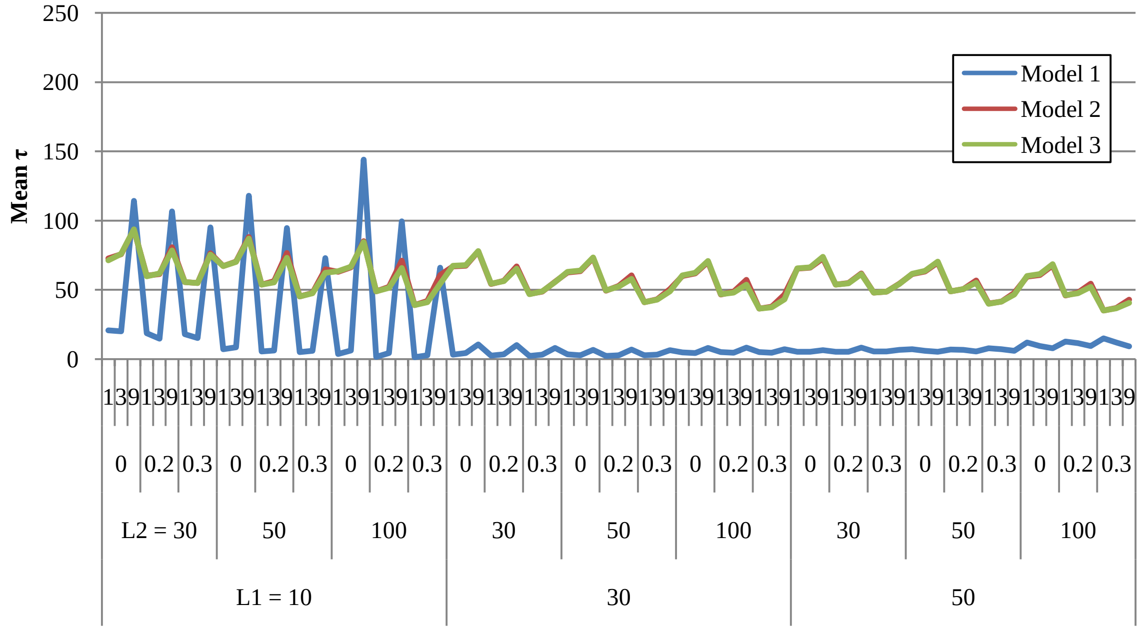


Figure 101: Mean τ_{11} of X_1 across Propensity Score Estimation Methods when using Between-Cluster Matching

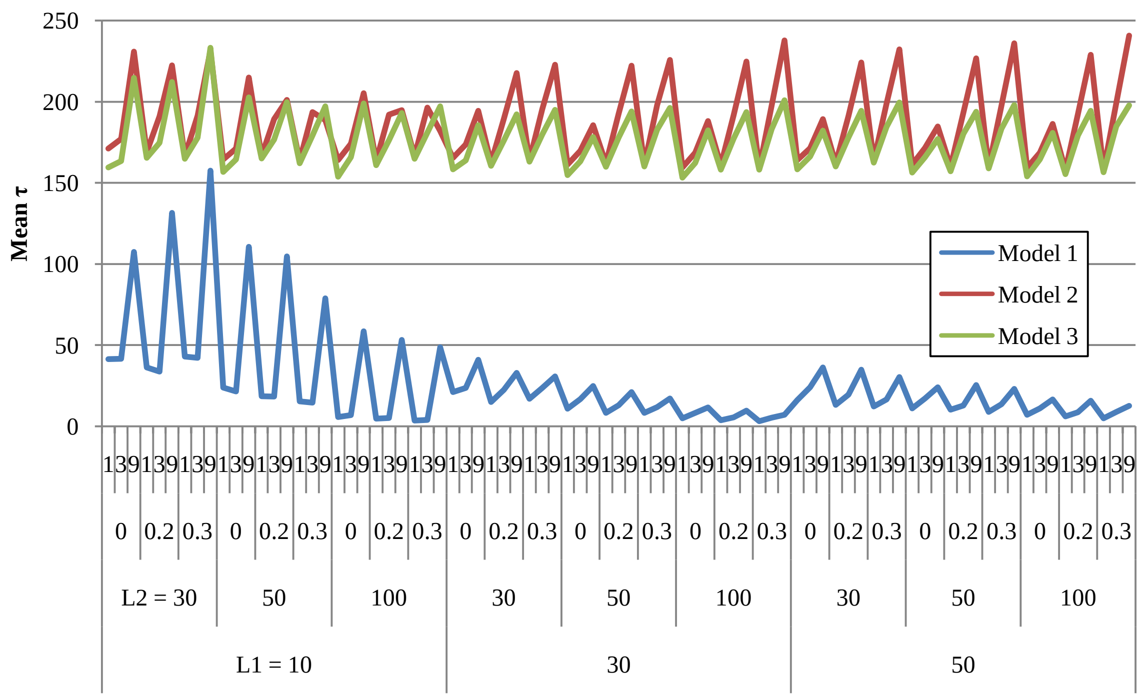
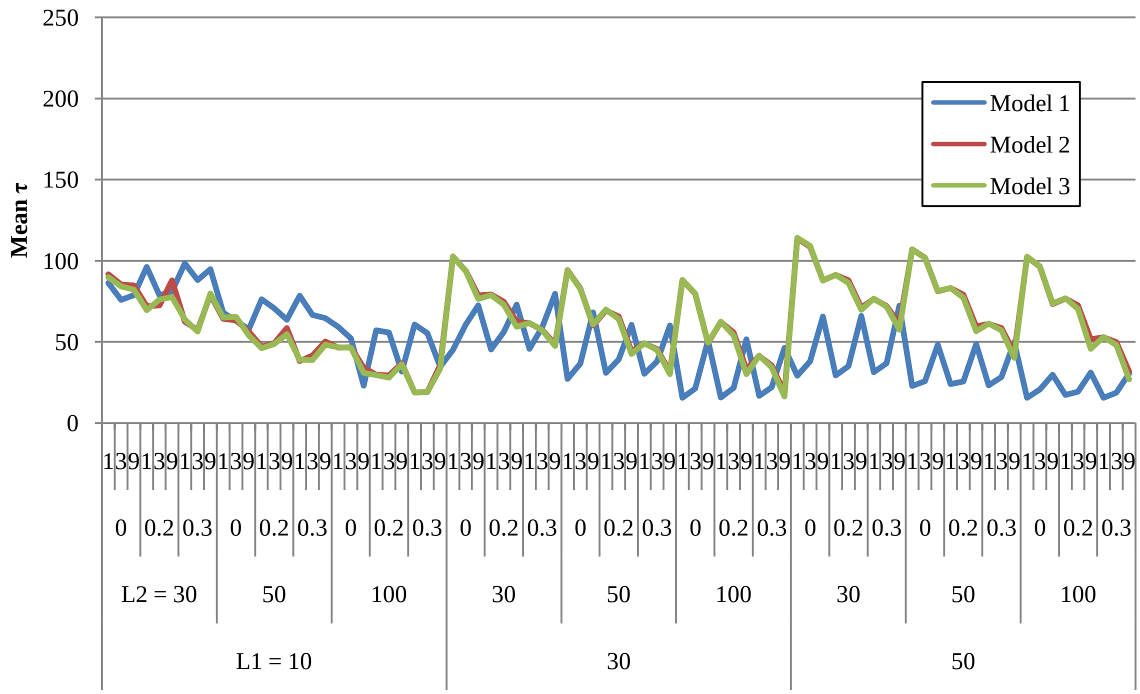


Figure 102: Mean τ_{11} of X_1 across Propensity Score Estimation Methods when using Quintile Stratification



Retention Rates of Initial Sample. The percent of the treatment group retained was smallest when balance was adjusted using within-cluster matching and largest when using quintile stratification. The percent retained that resulted from application of the propensity score estimated using Model 1 was at least half that of other models as a function of adjustment method.

Table 56:

Percent Treatment Group Retained across Propensity Scores and Adjustment Methods

Level-1 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
10	5.6	22.3	22.7	23.2	77.9	80.2	38.2	96.4	97.0
30	8.4	39.8	40.4	20.6	79.8	81.9	43.9	98.6	98.9
50	10.1	47.4	48.0	18.7	80.3	82.3	45.2	99.1	99.3
Mean	8.0	36.5	37.1	20.8	79.3	81.5	42.4	98.1	98.4

The retention of clusters was higher with the application of propensity scores estimated using logistic regression than propensity scores estimated through a multilevel model. The difference between retention rates at the cluster level per propensity score was not as great as that found between treatment group retention rates. Cluster retention rates were smallest across propensity scores when within-cluster matching was applied and greatest when quintile stratification was applied. The smaller retention rate resultant from within-cluster matching is likely caused by the requirement for close matches to occur within each cluster. If no close matches were available within the cluster, then the cluster would not appear in the matched sample. The retention rates resultant from between-cluster matching and quintile stratification were similar across propensity score estimation models, especially once sample size at level-1 reached 30.

Table 57:

Percent of Clusters Retained across Propensity Scores and Adjustment Methods

Level-1 Sample Size	Within-Cluster Matching			Between-Cluster Matching			Quintile Stratification		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
10	12.6	46.1	46.7	62.3	97.1	97.5	77.9	98.3	98.3
30	44.9	88.0	88.4	90.0	97.1	97.1	96.7	97.1	97.1
50	66.3	93.4	93.6	93.2	96.1	96.1	96.1	96.1	96.1
Mean	41.3	75.8	76.2	81.8	96.8	96.9	90.2	97.2	97.2

Applied Study

The performance of the propensity scores to balance covariates was explored using data made available by the National Center for Educational Statistics (2004) through the Educational Longitudinal Study of 2002. As in the simulation portion of this study, propensity scores were estimated using Models 1, 2 and 3. Each of these propensity score estimation models were applied to balance covariates using three methods: within-cluster matching, between-cluster matching, and quintile stratification. The context chosen for exploring balance achievement was through efforts to answer the question: Does family ownership of a computer have an effect upon math achievement scores? For this study, treatment group membership was defined as 10th-graders who do not have a computer in the home ($Z = 1$), and control group membership was defined as those who do have a computer in the home ($Z = 0$).

Variable Selection

As noted in the previous chapter, the decision to include a variable in the propensity score estimation model was based upon showing significant imbalance between the treatment and control group as well as evidencing a significant bivariate relationship with the outcome measure ($\alpha = .05$). The variables included in the propensity score estimation models are described in the following table. Descriptions of each variable are provided.

Table 58:

ELS:2002 Variable Description

Variable	ELS:2002	Description
Treatment Assignment		
Computer Ownership (Z)	BYS84C	Does family have a computer
Outcome Measure		
Math Score (Y)	F1TXMBIR	Follow-up Year, Math IRT number correct
Individual Level		
Math Score	BYTXMIRR	Base-line year, Math IRT number correct
SES	BYSES1QU	Socioeconomic status quartile
Computer Use for School Work	BYS45B	How often uses computer for school work
Computer Use on Own	BYS45C	How often uses computer to learn on own
Parent's Academic Expectation	BYP81	How far in school parent expects 10th-grader will go
Math Teacher's Academic Expectation	BYTM20	How far teacher expects student to get in school (MATH)

Students Academic Expectation	BYSTEXP	How far in school student thinks will get-composite
Family composition	BYFCOMP	Family composition variable: Biological mother and father, Two parent (non-biological), One parent, Other family structure
Math Sum	F1S17A - F1S17J	Total number of math courses
Math Breadth	F1S17A - F1S17J	Number of different of math courses
Race	F1RACE	African American, Caucasian, Latino/a
Cluster Level		
School Control	BYSCTRL	Public, Catholic, Other Private
% Free/Reduced Lunch	BY10FLP	Percent 10th grade free-reduced lunch

Propensity Score Estimation Models

The results of the propensity score estimation models are described in Table 59 below. For Model 1, which was estimated using a multilevel model, one variable showed significant variance across schools: Socioeconomic Status. Subsequently, only the impact of this variable upon the outcome was allowed to vary across schools in the estimation model. Those that resulted in parameter estimations that were found to be significant (α=.05) remained in the model. Two cross-level interactions were included in Model 1: The effect of the school variable, School Control, upon the relationship of SES and the treatment assignment; and the effect of the school variable, Percent Free/Reduced Lunch, upon the relationship between youth's use of computers to learn on his/her own and the

treatment assignment. Model 2 was estimated using logistic regression and included the two school-level predictors that were included in Model 1. Model 3 was estimated using logistic regression with no school-level predictors. No product terms were included in any of the estimation models. All parameter estimates reported in the table below are standardized in order to allow comparisons across variables with different scales.

Table 59:

Parameter Estimates Resultant from each Propensity Score Estimation Model

Effect	Parameter Estimate	Model 1	Model 2	Model 3
Intercept	γ_{00}	3.31*	5.15**	5.30**
Math number correct (BYTXMIRR)	γ_{10}	-1.53	-1.29	-1.20
Socioeconomic Status Quartile (BYSES1QU)	γ_{20}	-3.31**	-6.33**	-6.19**
Public	γ_{21}	-2.20*		
Catholic	γ_{22}	-0.93		
How often uses computer for school work (BYS45B)	γ_{30}	-1.43	-1.46	-1.41
How often uses computer to learn on own (BYS45C)	γ_{40}	-2.63**	-4.02**	-4.06**
Percent Free/Reduced Lunch	γ_{41}	-2.30*		
Parental academic expectation (BYP81)	γ_{50}	1.53	1.30	1.28
Teacher academic expectation	γ_{60}	-2.95**	-2.66**	-2.54*

(BYTM20)				
Two parents not both biological	γ_{70}	2.13*	2.09*	2.06*
One parent	γ_{80}	4.10**	3.48**	3.50**
Other family structure	γ_{90}	0.30	0.26	0.26
Student academic expectation	γ_{100}	-1.74	-1.04	-1.03
(BYSTEXP)				
Number of math courses	γ_{110}	-0.44	-0.63	-0.40
Breadth of math courses	γ_{120}	-0.73	-0.64	-0.84
Black	γ_{130}	1.40	0.61	0.55
White	γ_{140}	-3.03**	-2.88**	-2.84**
Latino	γ_{150}	2.11*	1.04	0.92
School Predictors				
Public	γ_{160}		-0.71	
Catholic	γ_{170}		-0.36	
Percent 10th grade free/reduced lunch	γ_{180}		-0.81	
(BY10FLP)				
Variance Components				
	τ_{11}	0.031	0.013	
	τ_{21}	-0.001	0.004	
	τ_{22}	0.002	0.001	

*significant at p<.05; **significant at p<.01

The direction of influence for each variable upon the treatment assignment was consistent across propensity score estimation models. For most variables, the level of significance was also consistent across models. An exception to this pattern was for the

variable “Latino” was not significant in Model 2 and Model 3 but was not significant in Model 1 ($p < .05$).

When interpreting the estimates, positive values indicate that an increase in the associated variable results in a greater likelihood of the youth not having a computer in the home. The results suggest that, as socioeconomic status increases, the likelihood not to have a computer decreases. A cross-level interaction was found between school-control and socioeconomic status. The school-control predictors were included in the model as dichotomous variables: Schools were identified as either being a public school or a Catholic school. The comparison group for these was Other Private Schools. The values for the cross-level interactions, therefore, suggest that the relationship of socioeconomic status with computer ownership is smaller in public schools than in private schools. The relationship between computer ownership and socioeconomic status did not significantly differ between students attending Catholic schools versus those attending other private schools.

As the amount of time that a youth uses a computer to learn on his/her own increases, the likelihood not to have a computer at home decreases. A significant and negative cross-level interaction was found to exist with this relationship and the variable indicating the percent of 10th-graders receiving free/reduced lunches in the school. The negative value for this predictor indicates that, as the percent of youth in the school who receive free or reduced lunches increases, the relationship between computer use and computer ownership decreases. Neither parental expectation of their child’s academic performance nor the youth’s own academic expectation was found to be significantly related to computer ownership; however, teacher’s academic expectation of the child was

related: As teacher expectation of youth's academic performance in math increased, the likelihood that the child did not own a computer decreased.

The family structure variables were entered into the propensity score estimation models as dichotomous variables, with the comparison group being those families with two biological parents. In this case, a family with two parents who were not both biological were more likely not to have a computer in the home than a family with two biological parents. The same relationship was found for those one-parent households, compared to two biological parent households, although the negative relationship was stronger.

The ethnicity variables were also entered as dummy-codes, with those who were White, Black, and Latino/a, each being compared to other ethnic groups. White youth were found to be significantly more likely to have a computer in the home than Other ethnic groups. Latino youth were found to be less likely to have a computer in the home compared to Other ethnic groups in the multilevel model, only. The likelihood of Black youth and the likelihood of Other youth to have a computer in the home did not differ significantly.

Balance Achievement: Mean γ_{10}

In order to assess the balance achieved for each student-level covariate between treatment and control group members as a result of the propensity score adjustment methods, a multilevel model was estimated for each covariate. As in the previous sections, each covariate was entered as an outcome of a multilevel model, and the treatment assignment was entered as the predictor (see Equation 2.5). In this model, both the intercept and treatment assignment were allowed to vary across schools. The values

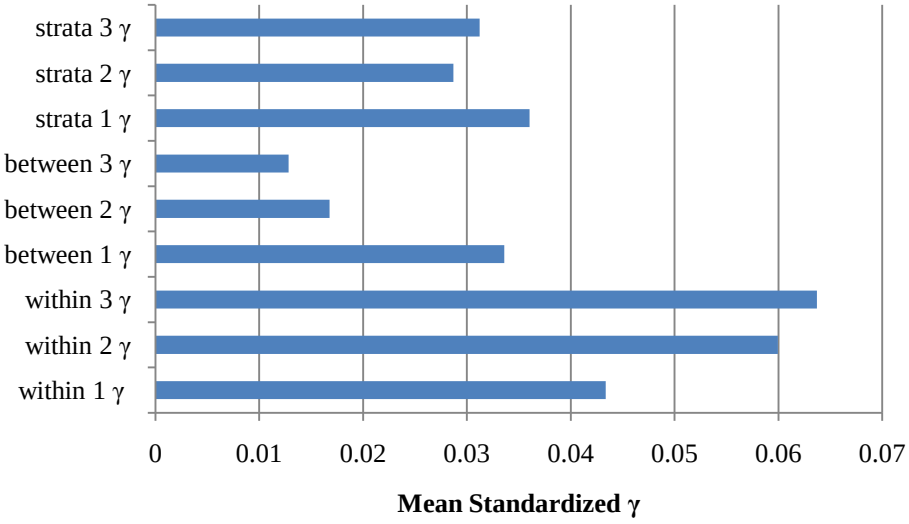
for γ_{10} were estimated and standardized. The mean standardized γ_{10} across the 16 student-level predictors included in the propensity score estimations are presented in Figure 103 below. It should be noted that all within-school matching methods resulted in too few individuals from family structures designated as "Other" to have γ_{10} estimated for that variable; therefore, the means for these methods include only 15 variables.

When using within-cluster matching to balance covariates, Model 1 showed the smallest mean γ_{10} when compared to balance from Model 2 and Model 3. When using between-cluster matching, the opposite pattern was apparent: Model 3 showed the smallest mean values of γ_{10} compared to balance from Model 2 and Model 1. When using quintile stratification, Model 2 showed the smallest mean values of γ_{10} compared to the balance achieved by Model 3 and Model 1. Across all adjustment methods and propensity score estimation models, between-cluster matching using Model 3 showed the smallest mean values for γ_{10} followed by between-cluster matching using Model 2 and quintile stratification using Model 2.

In order to determine the differences in performance between the different propensity score models and adjustment methods better, a two-way analysis of variance was performed. In this case, a 3 (method) × 3 (model) between-subjects factorial ANOVA was calculated comparing the standardized γ_{10} for the variables included in the model. The interaction effect between the method and model was not significant ($F(2, 135) = 2.032$, $p > .05$). A significant main effect for method was found ($F(2,135) = 13.117$, $p < .05$). Tukey's *HSD* was used to determine the location of the differences between the methods which revealed that the variables adjusted through within-cluster matching had higher values for γ_{10} ($M = .052$, $SD = .047$) than variables adjusted through

between-cluster matching (M = .021, SD = .023) and quintile stratification (M = .032, SD = .015). The main effect for model was not significant ($F(2,135) = .117$, $p > .05$).

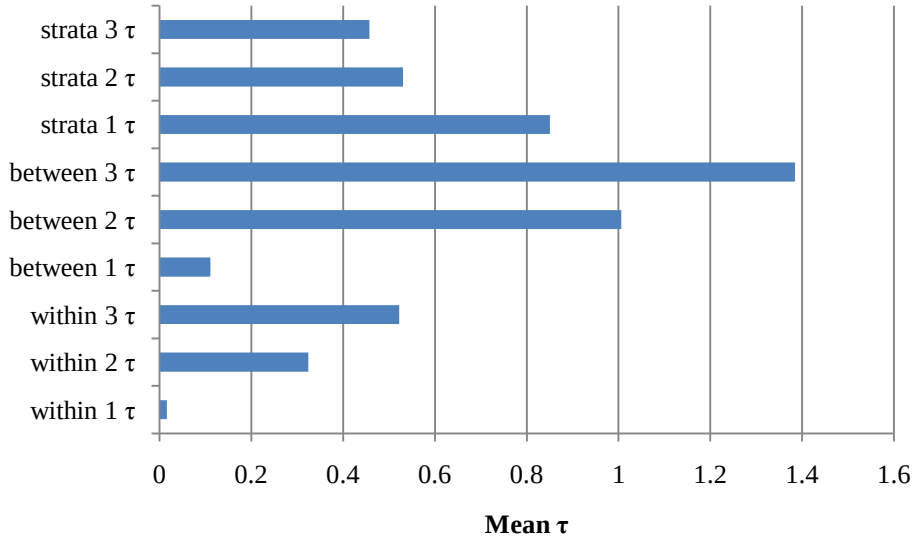
Figure 103: Mean Values for γ_{10} for All Predictors per Propensity Score Model and Adjustment Method



Balance Achievement: Mean τ_{11}

The value for τ_{11} provides an estimation of the variance in balance for each variable within each school. For both within-school matching and between-school matching, smaller values of τ_{11} were resultant from application of propensity scores estimated through Model 1 compared to both Models 2 and 3. Model 2 resulted in more consistent balance across clusters than did Model 3 when using matching techniques. A different pattern is apparent when the covariate imbalance is adjusted using quintile stratification: Propensity scores estimated using Model 1 show the greatest variance in the balance across schools, whereas Model 3 resulted in the smallest variance across schools.

Figure 104: Mean Values for τ_{11} for All Predictors per Propensity Score Model and Adjustment Method



When considering the variance in γ_{10} across schools for the two variables with significant cross-level interactions, all methods resulted in smaller values of τ_{11} for these variables than were apparent in the average τ_{11} across variables with one exception: For within-cluster matching with propensity scores estimated using Model 1, values for τ_{11} for BYS45C were greater than the average variance. This larger value for τ_{11} for this variable was still smaller than the mean τ_{11} for methods that did not use the propensity score estimated from Model 1.

Figure 105: Values for τ_{11} for Variables with Significant Cross-level Interactions

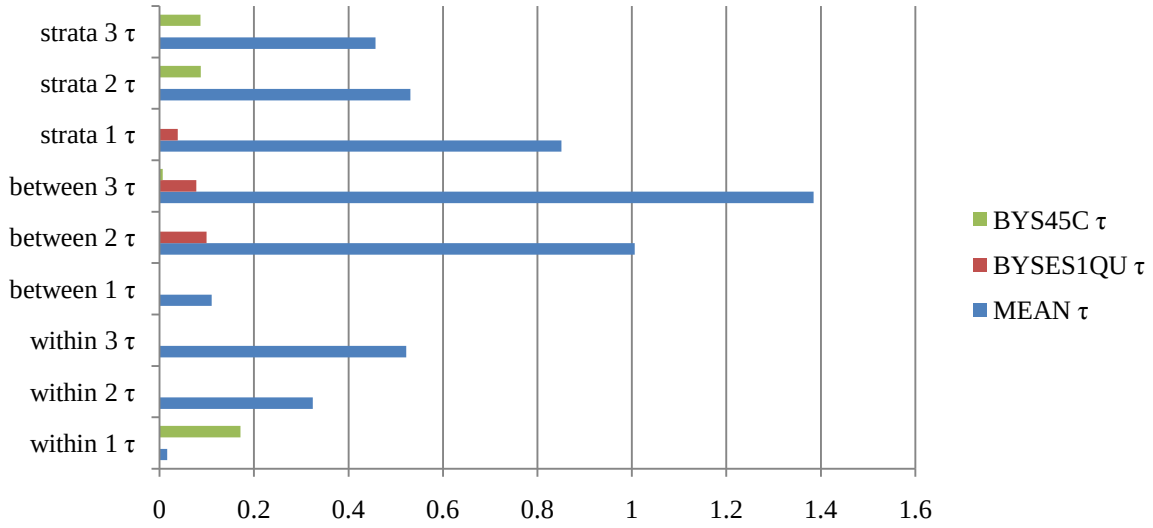
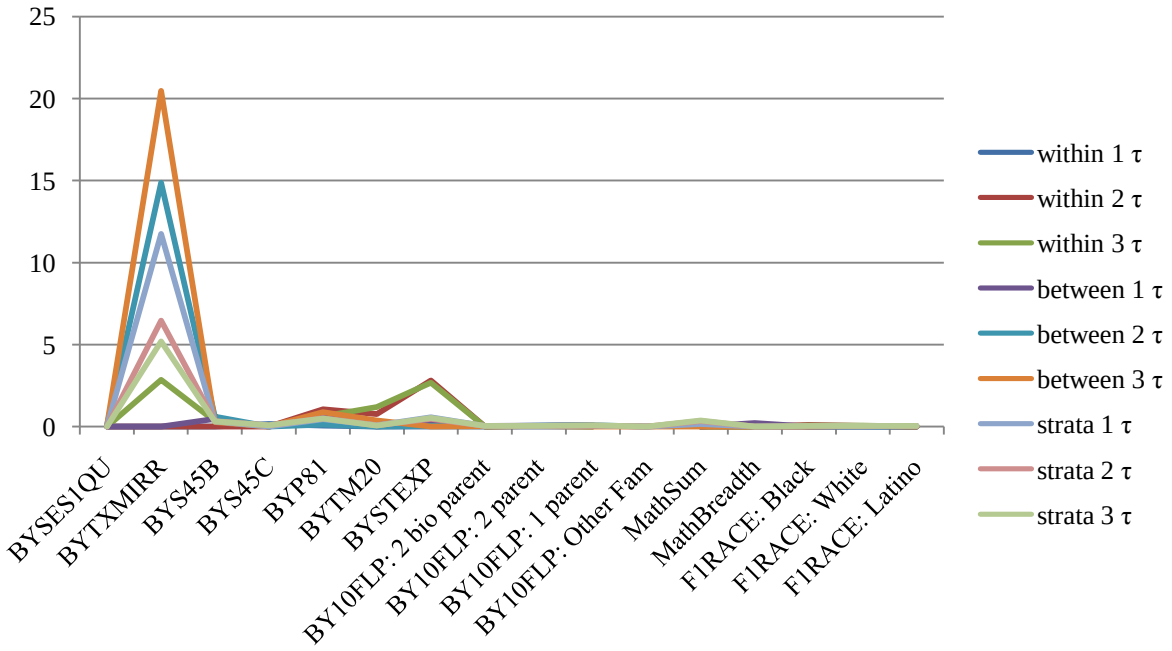


Figure 106 below illustrates that the mean value for τ_{11} for most methods was greatly influenced by the variance in balance across clusters for one variable: 10th grade math score (BYTXMIRR). Between-school matching using Models 3 and 2 showed the greatest variance in balance across clusters for this variable. The methods and models that showed the smallest values of τ_{11} for this variable were within-cluster matching using Models 1 and 2 and between-cluster matching using Model 1.

Figure 106: Values for τ_{11} for Each Variable across Propensity Score Models and Adjustment Methods

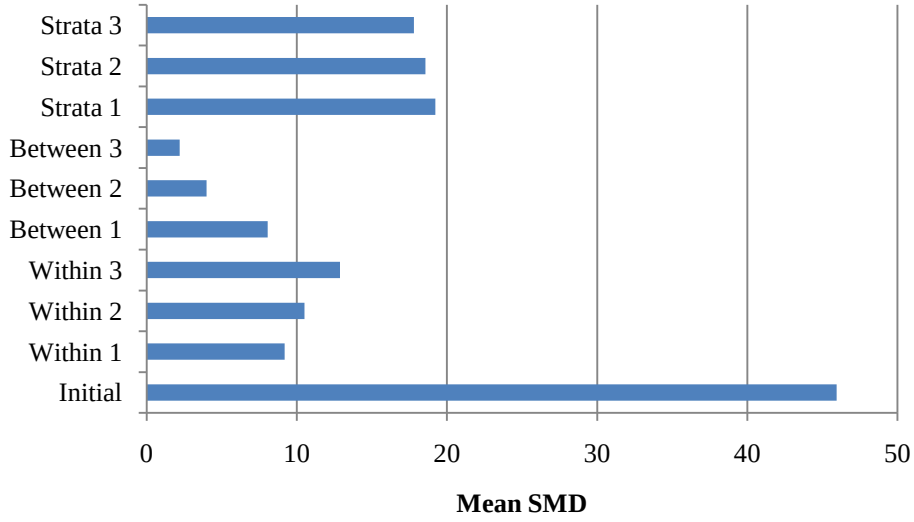


Standardized Mean Difference

The balance achieved per adjustment method and propensity score estimation model apparent in the standardized mean difference shows a similar pattern to that shown in the mean values of γ_{10}. An exception to these patterns is that within-cluster matching methods result in better balance than quintile stratification methods when SMD is the measure of balance. The opposite is true when mean γ_{10} is considered. These differences are likely related to challenges in estimating the values for γ_{10} when sample sizes are small, as is typically the case when within-cluster matching methods are applied. A second exception to the pattern between these two indicators of overall balance was found between the three propensity score estimation models in quintile stratification: In the case of SMD, the values were smallest for Model 3, whereas Model 2 showed the

smallest mean values of γ_{10}. Overall, all methods reduced the imbalance that was apparent in the initial sample.

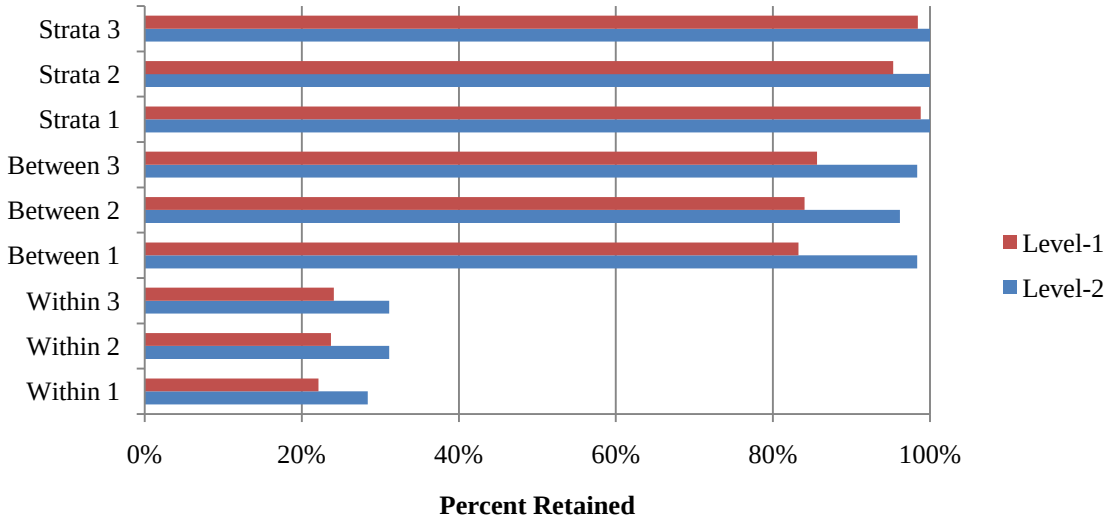
Figure 107: Mean SMD for All Predictors per Propensity Score Model and Adjustment Method



Retained Sample Size

Within-school matching resulted in much smaller retention rates of the treatment group members as compared to between-school matching and quintile stratification. This pattern was apparent in both level-1 and level-2 treatment group retention rates. Quintile stratification retained almost the entire treatment group and retained every school. Between-cluster matching showed retention rates slightly less than those with quintile stratification.

Figure 108: Percent Retained at Level-1 and Level-2



Bias in Treatment Effect Estimate

Because outcome data are available in the ELS:2002 data, the amount of bias resulting from the application of each adjustment method and propensity score estimation model can be estimated. The challenge is determining which method/model should be used as the comparison model. The method and model combination that resulted in the smallest values for overall balance was between-cluster matching using Model 3. This method/model combination did not, however, result in the smallest values for τ_{11}. The smallest values for τ_{11} resulted from application of Model 1 using within-cluster matching followed by Model 1 using between-cluster matching.

In order to assist in making the decision of which method should serve as the one to which others should be compared, the results from the simulation study were considered. The simulation conditions most similar to those apparent in the ELS:2002 data are presented in Table 60 below. No model/method was clearly the most successful at balancing covariates, although between-cluster matching and within-cluster matching

show similar performance for balance across the sample and within each cluster. Because between-cluster matching resulted in the retention of a much larger percentage of the treatment group members in the ELS:2002 sample (83% versus 22%) with little loss of balance, between-cluster matching using a propensity score estimated from Model 1 was selected as the comparison method/model.

Table 60:

Results from Simulation Study for a Level-1 Sample Size of 30, Level-2 Sample Size of 100, Mean Cross-Level Interaction of Zero, and a Treatment-Control Group Ratio of 1:3.

Method/Model	Mean SMD	Mean γ	Mean τ	Treatment Groups Retained
Within 1	7.90	2.09	4.60	8.5%
Within 2	26.95	0.09	61.74	42.3%
Within 2	31.33	0.11	62.29	42.6%
Between 1	5.06	3.80	8.48	17.7%
Between 2	34.71	1.95	168.46	84.0%
Between 3	35.0	2.02	162.31	85.0%
Stratif 1	9.98	2.47	21.24	45.2%
Stratif 2	84.79	1.18	79.68	100%
Stratif 3	83.11	1.26	79.73	99.4%

The treatment effect of family computer ownership upon math achievement was estimated using a multilevel model, allowing the intercept and the treatment assignment

to vary across schools. In the case of quintile stratification, the strata membership of each individual was dummy-coded and included in the model as a fixed variable. The initial estimate of the effect, as well as those resulting from each model by method combination, is presented in Table 61. As described in this table, the effect of computer ownership upon math achievement was found to be significant ($p < .05$) in the initial model and in when adjusting balance using quintile stratification with a propensity score estimated using Model 1. Once imbalances are addressed through the other adjustment methods/models, the effect of computer ownership is no longer significant.

Table 61:

Treatment Effect Estimates of Family Computer Ownership on Math Achievement

Adj. Method	Estimate	SE	df	t-value	p-value
Initial	-9.64	0.85	3879.00	-11.31	<.0001
Within 1	1.26	1.97	112.00	0.64	0.52
Within 2	-0.75	2.05	120.00	-0.36	0.72
Within 3	-0.07	2.07	122.00	-0.03	0.98
Between 1	-1.08	1.29	426.00	-0.84	0.40
Between 2	-0.32	1.27	430.00	-0.25	0.80
Between 3	-0.42	1.28	438.00	-0.33	0.74
Stratif 1	-1.78	0.89	1354.00	-2.01	0.04
Stratif 2	-1.41	0.88	1302.00	-1.60	0.11
Stratif 3	-1.46	0.88	1308.00	-1.67	0.10

The bias in treatment estimation that resulted from each method was calculated by calculating the difference between the estimate produced by the comparison Model/Method, which was between-cluster matching with a propensity score estimated using Model 1, and the estimate resulting from the given Model/Method combination. This difference is divided by the pooled variance of the two estimates. The resulting value was multiplied by 100. The results of the bias calculations are presented in Table 62 and sorted in order of increasing bias. All methods resulted in far less bias than was apparent in the initial sample. Interestingly, within-school matching using Model 1 showed very different results than those from between-school matching, although the indicators of balance were comparable. The primary difference was in the difference in number of treatment group members that the method was able to match successfully, within-cluster matching maintaining a much smaller sample size.

Table 62:

Bias Estimation

Method/Model	Distance from Between 1	Bias
Between 1	-	-
Within 2	0.34	19.71
Stratif 2	0.32	29.22
Stratif 3	0.38	34.33
Between 3	0.67	51.95
Within 3	1.02	59.02
Between 2	0.76	59.69
Stratif 1	0.70	63.17
Within 1	2.35	141.01
Initial	8.56	781.48

Implications for Educational Researchers

In order to offer some guidance to educational researchers who work with nested data and desire to adjust for covariate imbalances, the following summary is offered. It should be noted that these findings result from this singular study and additional exploration is recommended. In order to offer guidance regarding the choice of adjustment method or propensity score model, criteria for acceptable balance is designated. Assigning criteria on these measures is a subjective activity. Guidance has been offered through previous research regarding SMD which suggest that values that are

less than 10 indicate acceptable levels of balance (D'Agostino & Rubin, 2000; Rosenbaum & Rubin, 1985). The criteria selected for mean γ and mean τ are more subjective, and are selected in this case based upon the goal of attaining parameter estimates that are close to 0. A value of 0 for γ_{10} would indicate balance across the sample as a whole, and a value of 0 for τ_{11} would indicate approximately equal balance within each school. Based upon this consideration and considering the descriptive statistics for each indicator in the current study, values of γ_{10} that are less than 1 and values of τ_{11} that are less than 20 were selected as indicative of good balance. Finally, the percent of the treatment group that was retained per method was also considered, with values greater than 70% retained considered good. As noted in previous sections, treatment group retention rates do not indicate either good or bad performance in balancing covariates, but it is an important consideration when weighing the benefits and costs of selecting each method.

Under no simulation condition did a propensity score estimation model by adjustment method meet all criteria of covariate balance as described above. When the requirement of the 70% treatment group retention rate is removed, within-cluster matching and quintile stratification show some success when using propensity score estimation Model 1. This success was limited to the largest sample-sizes at level-1 (n=50), to conditions where the cross-level interactions were nonzero, and where the treatment-control group ratio was either 1:1 or 1:3. These findings support the idea that the application of propensity scores that are estimated using a multilevel model result in better covariate balance than those resulting from logistic regression models when the propensity scores can be successfully estimated. In order for propensity scores to be

successfully estimated, sample size at level 1 and level 2 should be large (e.g., at least 50 at both levels).

In addition to sample-size at level 1 and level 2, sufficient numbers of treatment group members should be present within each cluster. As illustrated in Table 63, the propensity score estimated using Model 1 did not meet criteria under any sample-size condition when the treatment-control ratio was 1:9, which is the condition in which treatment group membership is smallest. The smallest number of treatment group members per cluster that met criteria for successful covariate balance was 13. Additionally, the propensity score estimated using Model 1 did not meet criteria when the cross-level interaction was 0. A probable explanation for the poorer performance when the cross-level interaction was 0 is related to the inclusion of X_3, which served only to add noise to the estimation model rather than contribute to its accuracy. It is possible that, had X_3 not been included in the estimation model when the cross-level interaction was 0, the criteria for successful balance would have been met.

When sample size at level 1 was 50 and level 2 was 100, both quintile stratification and within-cluster matching methods successfully met these three criteria. Given that the percentage of the treatment group retained for quintile stratification was approximately five times that resulting from within-cluster matching for the same simulation conditions, quintile stratification shows clear advantages over within-cluster matching.

Table 63:

Successful Balance Achievement as Indicated by SMD, Mean γ, and Mean τ

Simulation Conditions				Within-Cluster Matching	Quintile Stratification
L1	L2	$\rho_{(WX)Z}$	Ratio	using Estimation Model 1	using Estimation Model 1
50	50	.2	1:1	X	
50	50	.2	1:3	X	
50	50	.3	1:1	X	
50	100	.2	1:1	X	X
50	100	.2	1:3	X	X
50	100	.3	1:1	X	X
50	100	.3	1:3	X	X

Because so few conditions met these three criteria, those that showed balance as indicated by the mean γ and mean τ criteria, only, were also explored. When both the SMD and treatment group retention rate were dropped as a qualifying criteria, the success of the within-cluster matching method becomes apparent. Within-cluster matching using propensity scores estimated using Models 2 and 3 showed successful balance in a number of the small sample-size conditions where other methods did not meet criteria, with Model 2 appearing in more conditions than Model 3. Models 2 and 3 also showed success in balancing covariates in the 1:9 ratio condition where results from Model 1 did not. The success of Model 2 and Model 3 over Model 1 in the smaller sample size conditions is likely related to success in estimating propensity scores. Estimation models that used logistic regression resulted in balance with as few as 1 treatment group member per

cluster (i.e., L1=10 with a 1:9 ratio), although this required 50 clusters. Including the cluster-level predictors in the model (Model 2) resulted in a greater number of conditions meeting criteria of balance than when the cluster-level predictors were not included (Model 3). Also important to consider are those models and methods that do not appear in Table 63 and Table 64. The exclusion of quintile stratification methods that used Models 2 and 3 and all between-cluster matching methods was related to the τ_{11} criterion. Between-cluster matching and quintile stratification methods resulted in larger values for τ_{11} than did within-cluster matching methods under most simulation conditions.

Table 64:

Successful Balance Achievement as Indicated by Mean γ and Mean τ

Simulation Conditions				Within-Cluster Matching			Quintile
L1	L2	$\rho_{(WX)Z}$	Ratio	Model 1	Model2	Model3	Stratification using Model 1
10	30	.3	1:3		X		
10	50	0	1:9		X		
10	50	.2	1:1		X	X	
10	50	.2	1:3		X		
10	50	.2	1:9		X		
10	50	.3	1:1		X	X	
10	50	.3	1:3		X	X	
10	50	.3	1:9		X		
10	100	0	1:1		X	X	
10	100	0	1:3		X		
10	100	0	1:9		X	X	
10	100	.2	1:1		X	X	
10	100	.2	1:3		X	X	
10	100	.2	1:9		X	X	
10	100	.3	1:1		X	X	
10	100	.3	1:3		X	X	
10	100	.3	1:9		X	X	
30	30	.3	1:3		X		

Simulation Conditions				Within-Cluster Matching			Quintile
L1	L2	$\rho_{(WX)Z}$	Ratio	Model 1	Model2	Model3	Stratification using Model 1
30	30	.3	1:9		X		
30	50	.2	1:1	X			
30	50	.2	1:3	X			
30	50	.2	1:9		X		
30	50	.3	1:1	X	X		
30	50	.3	1:3	X	X	X	
30	50	.3	1:9		X	X	
30	100	0	1:9		X		
30	100	.2	1:3		X		
30	100	.2	1:9		X	X	
30	100	.3	1:1		X	X	
30	100	.3	1:3		X	X	
30	100	.3	1:9		X	X	
50	50	.2	1:1	X			
50	50	.2	1:3	X			X
50	50	.3	1:1	X			
50	50	.3	1:3		X		X
50	50	.3	1:9		X		
50	100	0	1:9				X
50	100	.2	1:1	X			X

Simulation Conditions				Within-Cluster Matching			Quintile Stratification using Model 1
L1	L2	$\rho_{(WX)Z}$	Ratio	Model 1	Model2	Model3	
50	100	.2	1:3	X			X
50	100	.2	1:9		X		X
50	100	.3	1:1	X	X		X
50	100	.3	1:3	X	X		X
50	100	.3	1:9		X	X	

Finally, the current study offers a few suggestions to researchers regarding optimal sample characteristics. A common question that researchers face is whether it is more important to add more individuals to each cluster or to add more clusters. The findings from this study suggest that, when applying propensity score methods to nested data, it is most important to have a sufficient number of individuals within each cluster from which to efficiently estimate the cluster-level effects. When estimating a propensity score using a multilevel model, there is limited benefit to adding clusters above 30. Benefits to balance are apparent, however, when adding individuals to a cluster, with a level-1 sample size of 50 resulting in better overall balance than a level-1 sample size of 30 or 10. Findings also indicate that it is generally better to select a more balanced ratio of treatment and control group members with which to match than collecting many control group members. This is especially relevant when using quintile stratification when applying propensity scores that are estimated using logistic regression.

Future Research

The opportunities for research abound in the area of propensity score adjustment methods using clustered data. Because propensity scores that are estimated using a multilevel model showed particular sensitivity to X_3, which had the smallest relationship with the treatment assignment, greater exploration of extraneous variables and their effects on propensity score estimations would be valuable to researchers who work with clustered-data. Additionally, the noise that X_3 added to the propensity score estimations might have been amplified because X_1, X_2, and X_3 were perfectly uncorrelated in the generating sample, a situation which rarely occurs with observational data. Had these variables been correlated, the propensity score adjustment methods would likely have met criteria of balance under a greater number of simulation conditions, but further research is needed to explore such a hypothesis.

Another characteristic of the samples resultant from the generating model that is uncommon in observational data is related to the cross-level interactions. Cross-level interactions often exert influence on the treatment assignment mechanism through only a few variables, rather than equally across all variables. Exploring the effects of the cross-level interaction upon a single covariate used in estimating a propensity score would be a valuable contribution to educational research.

The generating models for the current study established a relationship between each X and the treatment assignment, which can be expressed as a slope with a variance. The ICCs of those slopes was set to 0.1, which created different slopes within each cluster, but those slopes ultimately averaged to the slope designated in the generating model for that X. The effect of this characteristic of the generating model upon the

estimation and performance of the propensity scores would be of interest to explore. Changing this variance in the slope would be particularly interesting for X_3, which had an average slope of 0 across clusters. If τ_{33} were set to 0, then X_3 would be a true nuisance variable in the propensity score estimation models. Values of τ_{33} that are nonzero, however, would result in X_3 contributing to the estimation of the treatment assignment mechanism. Better understanding of the influence of X_3 given different variance conditions would better inform researchers as to the effects of propensity score methods for adjusting for imbalances in nested data.

References

Austin, P. C. (2008). The performance of different propensity-score methods for estimating relative risk. *Journal of Clinical Epidemiology, 61*, 537-545.

Austin, P. C., Grootendorst, P., Normand, S. L. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine, 26*, 754-768.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31072.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-314.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankya, Ser. A, 35*, 417-446.

D'Agostino, R. B., Jr., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association, 95* (451), 749-759.

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

Du, J., Havard, B., Yu, C., & Adams, J. (2004). The impact of technology use on low-income and minority students' academic achievement: Educational longitudinal study of 2002. *Journal of Educational Research and Policy Studies, 4*(2), 21-38.

Estes, K. (2008). *Sample Size in HLM*. Unpublished doctoral dissertation. Georgia State University.

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

Hong, G. (2004). Causal inference for multilevel observational data with application to kindergarten retention. Unpublished doctoral dissertation. University of Michigan.

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

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

Hox, J., & Maas, C. (2006). Multilevel Models for Multimethod Measurements. *Handbook of multimethod measurement in psychology* (pp. 269-281). Washington, D.C.: American Psychological Association.

Heckman, J., Ichimura, H., & Todd, P. (1998). Matching as an economic evaluation indicator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605-654.

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrics, 66*(5), 1017-1098.

Heckman, J., Ichimura, H., & Todd, P. (1998). Matching as an economic evaluation indicator. *Review of Economic Studies, 65*(2), 261-294.

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

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706-710.

Joffe, M. M. & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology, 150*(4), 327 – 333.

Kim, J. & Seltzer, M. (2007). Causal inference in multilevel settings in which selection processes vary across schools. CSE Technical Report 708. University of California.

Lu, B., Zanutto, E., Hornik, R., & Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association, 96*(456). 1245-1253.

National Center for Educational Statistics (2004). *Educational longitudinal study: 2002 data files and electronic codebook system* (NCES 2004–004). [CD-ROM]. Washington, DC: Author.

Oakes, J. M. & Johnson, P. J. (2006). Propensity score matching for social epidemiology. In J. M. Oakes & J. S. Kaufman (Eds.), *Experimental social epidemiology* (pp. 370-392). San Francisco, CA: Jossey-Bass.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd. Ed.).* Thousand Oaks, CA: Sage.

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

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.

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

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

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2*(3), 808-840.

Rubin, D. B. (1997). Estimating causal effects from large scale data sets using propensity scores. *Annals of Internal Medicine, 127*, 757-763.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

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

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143-157). Washington, D.C.: American Psychological Association.