Improving and Assessing Propensity Score Based Causal Inferences in Multilevel and Nonlinear Settings

by

Benjamin M. Kelcey

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Education)
in The University of Michigan
2009

Doctoral Committee:

Professor Kenneth Frank, Co-Chair
Professor Brian P. McCall, Co-Chair
Professor Joanne F. Carlisle
Assistant Professor Bendek B. Hansen

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Improving and Assessing Propensity Score Based Causal Inferences in
Multilevel and Nonlinear Settings

by

Benjamin M. Kelcey

Co-Chairs: Kenneth Frank and Brian McCall

Recent calls for accountability have focused on scientifically based research that isolates causal mechanisms to inform both the policies and practices of education. A major challenge in aligning educational research with such standards has been to develop methods that can address the interdependency and multilevel structure of teaching and learning and approximate randomized experiments using observational data. In this dissertation, I carried out three studies that centered on improving causal inferences drawn from observational studies in common educational settings. In the first study, I developed several models for estimating multilevel propensity scores (PSs) and examined their effectiveness for causal inference. The results suggested consistent gains from multilevel PSs that allow differential influence of the group on its individuals. The results further suggested that covariate selection in multilevel PSs can play a large role, both relative to model type and in an absolute sense. The second study then developed a method to construct PSs in an effective and efficient manner using two pivotal relationships. The method made use of each covariate's relationship with the treatment and commonly available outcome proxies (e.g. pretest measures) to construct PSs that minimizes the mean-square error (MSE) of the treatment effect estimator. The results of the study suggested that an effective and efficient approach to constructing the PS might be to include those covariates whose relationship with the outcome is at least half the magnitude of the respective relationship with the treatment. In the final study, I develop an index that assesses the sensitivity of inferences in binomial regression models by extending the impact threshold of a confounding variable framework (Frank, 2000). Each of these methods is then applied to observational data to demonstrate how these methods can advance the quality and robustness of causal inferences in educational research.

CHAPTER I

Introduction

Educators, educational policy makers and educational researchers must attend to causal inference because education is fundamentally a pragmatic enterprise (Raudenbush, 2005; Frank, Maurolis, Duong & Kelcey, 2009). If a new program, pedagogy or school structure contributes to learning, then we should adopt it (Cook, 2002). If not, then we should not. With the current environment of accountability and passage of federal educational programs such as No Child Left Behind (NCLB), attention has focused on the need for evidenced-based educational research. The pragmatic nature of education gives rise to a need for discerning the causal effects of various educational interventions. The empirical evidence generated through such studies in education has pervasive policy implications that are useful at both local and national level. In recent years, such empirical evidence has been aimed at 'scientifically based research'. In particular, Part A., Sec. 9101 of the No Child Left Behind Act, defines "Scientifically Based Research" as

> (A) …research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs; and
>
> (B) includes research that ...
>
> (iv) is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs

> to the extent that those designs contain within condition or across-condition controls;..."

In research, this focus on scientifically based educational research has predominantly translated into research studies based on the principles of randomized experimental designs as they tend to balance both observed and unobserved pretreatment characteristics of experimental units.

Experiments potentially offer more robust and unbiased estimates of the treatment effect because the expectation is that they balance both measured and unmeasured pretreatment differences. This property tends to reduce the threat of imbalanced group assignment. However, the balancing property is often mitigated by several practical considerations. One such factor is the difficulty of implementing a truly randomized experiment. Among other concerns, it can be difficult to ensure that subjects assigned to the treatment group were actually assigned at random. One can envision an experiment where the treatment is tutoring for students in an after school program. Though in an experiment this additional resource would be randomly assigned, in actual practice teachers or principals might assign those students that they feel need the tutoring program the most. Another factor is the high cost of an experiment. As the balancing properties of an experiment are asymptotic, it is important that sample sizes for groups are relatively large and align with the hypothesized treatment effect size. Attention to such considerations in educational interventions enacted at the teacher or school level are particularly pertinent as the relevant sampling units are schools or teachers rather than students. From a more philosophical perspective, it has also been suggested that experiments represent a certain departure from reality (Heckman, 2005). That is, in authentic representations of phenomena, subjects self-select into or are placed by program organizers in treatment groups that are presumably beneficial to them (Heckman, 2005). In an experiment people are artificially placed in a treatment condition they would not necessarily choose to be part of. This argument is similar in nature to the medical effectiveness versus efficacy consideration. Whereas the medical effectiveness of a drug can be viewed as the impact of a drug under near perfect conditions, the efficacy of a drug is the impact it has under normal widespread use.

The difficulty of implementing randomized studies in education has historically made observational studies more feasible when considering the effects of treatments or policies that are difficult to randomize. In this type of study, where there is an absence of random assignment, groups are potentially imbalanced on pretreatment characteristics that may or may not influence the outcome. That is, there may be systematic differences between the treatment groups beyond the treatment status. As a result, differences in outcomes may be a function of pretreatment imbalances between the groups rather than a treatment effect. A historic example of this may be found in assessing the benefit of attending a private Catholic school. That is, researchers have asked whether attending a private Catholic school has a significant impact on students' learning above and beyond the impact seen in public schools. Though private Catholic school students frequently score higher on common achievement tests, opposing arguments have suggested that this difference in achievement is a result of private Catholic school students maintaining a higher aptitude or supportive family to begin with. In other words, the opposition has suggested that the population of students in public schools is dissimilar from the population in private Catholic schools. To address such systematic differences, researchers have developed a number of tools to adjust for differences using analytic methods such as matching, stratification or model based covariance adjustments. Consequently, a primary objective in the design and analysis of observational data is to control for alternative explanations through statistical adjustment. Although many such adjustments offer accurate estimation of treatment effects, their accuracy is tempered by the threat of an unmeasured baseline imbalance between the groups. Further, it is likely unrealistic in a variety of situations to assume that every factor that influences the treatment assignment has been measured. In such a case, and potentially in every observational study, two subjects who are identical on measured characteristics may have an unequal probability of being assigned to the treatment or control groups because of differences on an unmeasured characteristic. Estimates of the treatment effect when this residual bias is present can not be considered causal estimates as the relationship can plausibly be attributed to alternative differences between groups. Though inferences surrounding causal effects are generally more reliable when they are based on randomized experiments, observational studies are commonly used for answering causal questions in education for a number of reasons alluded to earlier. Experiments in education, for example, tend to have a high cost compared to observational studies. Further such randomized studies are generally less representative of the target population than observational studies. Another reason may be related to theory and hypothesis building. Observational studies are often more apt to answer a number of exploratory questions. They are frequently well suited to provide a reasonable basis for preliminary assessment of a treatment and subsequent design of a randomized experiment. Yet another reason is the complexity of the education process. For instance, in assessing the effect of teacher knowledge on a student’s achievement it is difficult to envision how one might conduct a randomized experiment. That is, because teacher knowledge is likely a characteristic acquired and developed over a large span of time, it is extremely difficult treatment to randomize. However, given the current era of school and teacher accountability and emphasis on teacher quality, one might suspect that teacher knowledge might improve both teacher quality and the learning of students. Given such impetus, how might we assess the contribution of teacher knowledge to student achievement? Though randomized experiments would be optimal in most situations, they are likely difficult in the current example (e.g. Rubin, 1974). As a result, researchers often rely on observational data and analytical adjustments to assess the effects of complex treatments. A central goal then is to improve the quality of inferences drawn from observational data.

The purpose of this dissertation was to improve the value of such causal inferences in observational studies. In particular, the dissertation examined methods to improve causal inferences drawn from observational studies along three general dimensions of causal inference. These three general dimensions are the design of the study, the analysis of the data and drawing inferences from such analyses (Rubin, 2007). More specifically, the design dimension underscores the intended plan to setup treatment groups and measure important variables for a sample of the target population such that useful inferences may be drawn from the study. The analytic dimension emphasizes the use of statistical adjustments to address imbalances among treatment groups to serve two primary purposes: to increase the precision of comparisons and to remove initial bias due to confounding variables. The final dimension, drawing inferences, draws attention to the generalizability of the analytic findings and speculates how hidden sources of bias may alter such inferences. Along these three dimensions, I attempted to improve causal inferences using observational data in several capacities pertinent to education.

The first study was aimed at advancing the capacity of observational studies in terms of study design as well as analysis. In particular, I attempted to improve such inferences by developing several approaches to estimate treatment effects using a more holistic view of how treatments are assigned in educational settings. Teaching and learning represents a multilevel system where students, teachers and schools affect each other and do so in differential manners. As a result, a pivotal consideration in assessing the effects of treatments within such schooling systems is the reciprocal influence of schools, teachers and students. To attend to this influence, the first study developed and assessed several different multilevel propensity score (PS) structures one might consider in adjusting for group influence in treatment assignment. That is, the first study attempted to improve inferences in these systems by explicating the processes by which groups (e.g. schools) may influence individual treatment assignments. I conceptually developed potential treatment assignment mechanisms and adapted PS methods to address the manners in which groups might help decide the treatment status of an individual. Such developments improve causal inferences by identifying not only comparable individuals, but also individuals who are in comparable groups and are influenced by their respective groups in similar manners. Consequently, inferences are now based on the contrasting outcomes among those individuals who have similar personal and group characteristics rather than just similar personal characteristics. Further, the study attempted to explicate the comparative contribution of PS model type to variable specification in terms of the quality of the treatment effect estimator. More specifically, I examined several PS model types and the inclusion of different covariates in estimating the treatment effect in terms of bias, variance and mean-squared error (MSE). Such exploration attempted to make clear the influence of different model and variable types so as to inform PS construction in multilevel settings. More generally, such exploration attempted to inform the study design dimension by assessing the need to measure and control for groups' contribution to overt bias. Further, the study addresses the analytic dimension by contrasting the different analytical models one might use to adjust for the group influence in treatment assignment. The method was then applied to a study concerning the effect of teacher's literacy knowledge on first graders' reading achievement.

The second study was positioned to improve the analytic dimension as well as inform the design dimension of causal inferences. The second study improved inferences drawn from observational studies by improving the quality of the treatment effect estimators. In particular, by making use of measured variables, I attempted to improve causal inferences by identifying the most effective and efficient variables to include in the PS. Though it is common in practice to include all available variables when specifying the PS and/or to model the treatment assignment ignoring each variable's relationship with the outcome, recent analytical and empirical results have suggested otherwise (e.g. Brookhart et al., 2006). In particular, research has shown that the reduction in bias by including a variable in the PS can be exceeded by a decrease in efficiency. Further, ignoring variables' relationship with the outcome discounts the duality of confounding by constructing comparison groups that are not the most effective and efficient for a given outcome. That is, the influence of a variable on the bias or variance of an estimator depends on a variable's relationship with both the treatment and the outcome. For instance, including or excluding a variable strongly related to the treatment but unrelated to the outcome does not bias the estimator, however including it does add substantial variance to the estimator. By linking two pivotal covariate relationships with the MSE of the treatment effect estimator, I estimated thresholds at which a variable is likely to increase or decrease the MSE of the treatment effect estimator if included in the PS. More conceptually, the background variables one attempts to match individuals on tend to play a large role in the accuracy and consistency of the treatment effect estimate. As a result, I developed a method to construct the PS such that the bias and variance of the treatment effect estimator are jointly minimized given the observed data. Such construction attempts to identify those variables whose reduction in bias is exceeded by their contribution to variance. Accordingly, the treatment effect estimator tends to have a density centered at and concentrated around the true treatment effect. Inferences from observational data are improved by selecting the covariates that provide a combination of a precise and accurate estimate given the data. More generally, the study informs the analytic dimension of causal inference by identifying those variables to include in the PS and informs the design dimension by identifying those variables whose measurement is pivotal. The study is followed by an example in which I examined the effect of kindergarten retention as a school policy on the school's average math achievement.

In the final study, I attempted to inform the debate about inferences drawn from observational studies by assessing the robustness of such inferences. This study was primarily intended to advance the inferential dimension of causal inference. Whereas the first two studies advanced the design and analysis dimensions of inference drawn from observational studies, they did not address the threat of an unobserved confounding variable. Because observational studies do not randomize the treatment assignments, the potential for unobserved variables to be imbalanced among treatment groups always exists. As a result, though a primary objective of causal inference in observational studies may be to identify, measure and adjust for confounding variables, a secondary objective is to then speculate about the remaining unobserved bias. To assess the unobserved bias or imbalance needed to change such inferences, I developed a framework to estimate the magnitude of an imbalance needed to change an inference. More specifically, I made use of two pivotal covariate relationships to extend the Impact Threshold of a Confounding Variable to binomial regression models (BRMs) (Frank, 2000). That is, similar to a sensitivity analysis, I developed thresholds that index the minimum relationships an unobserved covariate must have to invalidate an inference. The method I developed allows one to quantify the absolute sensitivity of an inference in BRMs by ascertaining a specific threshold at which an inference is invalidated. The results inform causal inference by quantifying the conditions necessary to invalidate a statistical inference at a given α level. In turn, I applied this index to an international data set (SACMEQ) concerning the effect of parental education level on reading achievement.

Throughout each of these studies and settings, I additionally focused on the roles of specific covariate relationships in informing causal inference. Because a primary task of observational studies is to measure and adjust for confounding variables, I studied the influence of different variable relationships on the properties of the treatment effect estimator. In particular, I examined the role of each covariate's relationship with the outcome and the treatment and how they might guide and inform inferences as well as study design in multilevel and nonlinear settings. Throughout the three studies, a common condition and implication emerges: in designing a study, estimating an effect and drawing inferences, the variables measured and adjusted for tend to dominate the model or structure in which they are considered. As a result, a second common thread among the studies is assessing and understanding the differential contributions of variables and models to the quality of the treatment effect estimator. In each of the studies, focus is additionally placed on harnessing the influence of covariates to improve estimation. That is, by linking each covariate's unique relationship with the outcome and treatment to the treatment effect estimator, I developed methods to improve the properties of the treatment effect estimator. In particular, in my first study I investigated the contribution of covariates with different relationships among several multilevel PS structures in terms of the effectiveness and efficiency of the treatment effect estimator in multilevel settings. The second study then examined how such covariate relationships can be used to improve treatment effect estimators in multilevel settings when using the PS. Finally, in my third study I used those same covariate relationships to understand the robustness of inferences in non-linear settings by extending the impact threshold of a confounding variable to binomial regression models (Frank, 2000). Using this framework and line of inquiry, this dissertation attempted to advance causal inference in education observation data via three statistical methods.

CHAPTER II

The Roles of Multilevel Propensity Scores and Variable Selection in Multilevel Settings

Introduction

Causal inference in educational settings often poses unique complexities as a result of the multilevel structure in which teaching and learning take place. In particular, students and their learning experiences tend to be grouped within a class and in turn such classes are clustered within schools. Such grouping or clustering often invokes dependencies among students within the same class and school. For instance, a disruptive student in one class may affect the learning opportunities in another. In addition, students within a classroom share the same teacher and his or her ability to teach effectively. Further, students within a school tend to share similar resources and are governed by the same management and policies. In estimating conditional associations between outcomes and treatments of interest, such dependencies within the data have been considered through a variety of structures such as multilevel modeling (e.g. Raudenbush & Bryk, 2002).

One such framework that attempts to advance such associational inference to causal inference is the Rubin Causal Model (RCM) (e.g. Holland 1986). As previously discussed, the RCM formalizes causal inference by focusing on the idea of potential outcomes and a treatment assignment mechanism. In particular, under the RCM and its additional assumptions, unbiased estimates of the treatment effect can be obtained if the treatment assignment is conditionally independent of the potential outcomes given the observed covariates. One method that attends to such a framework by providing such conditional independence or strong ignorability of the treatment assignment is the propensity score (PS) (Rosenbaum & Rubin, 1983a).

Unlike standard parametric methods that control for confounding in an outcome model, PS methods rely on a model of treatment assignment to adjust for confounding (Brookhart, Schneeweiss, Rothman, Glynn, Avorn & Sturmer, 2006). In observational studies where researchers do not know the true treatment assignment mechanism, researchers attempt to infer the treatment assignment mechanism from the observed data. Though literature has addressed multiple methods to infer such assignment, literature has been relatively scarce concerning those structures common in educational studies. In particular, the multilevel structure of educational data presents properties atypical to common applications of the PS. In other words, the nested structure in observational studies in education makes inferring the treatment assignment mechanism more difficult. For example, in observational education studies that resemble multi-site randomized trials, treatments are assigned to individuals within each site. Such assignment is often based on a student's characteristics as well as his/her school's characteristics and membership. Correspondingly, selection bias may originate from both the individual and group level. When inferring the treatment assignment mechanism in such studies one must consider the contribution of individuals to their selection into treatments but also the contribution of their respective groups to their treatment assignments. That is, not only are the outcomes multilevel, but also the treatment assignment mechanism is multilevel.

Thus, a central issue in extending PS methods to observational education data is to appropriately address the potential nested nature of treatment assignment. Further, inferring such multilevel mechanisms can be difficult in the presence of many pretreatment covariates that may plausibly influence the treatment assignment. In particular, as the influence of the group on the individual's treatment assignment grows, researchers must not only consider the pretreatment covariates but also the interactions between groups and individuals. As the inclusion or exclusion of such pretreatment covariates can strongly affect the subsequent bias and variance of the treatment effect estimator, a second issue facing researchers using multilevel PS methods is how to select variables to be included in the PS model.

Literature has proposed hierarchical generalized linear models (HGLMs) with random intercepts to specify an individual's treatment probability (Hong & Raudenbush, 2006; Kim & Seltzer, 2007). This research has suggested HGLMs were superior to single level models as they produced better balance on covariates. Yet the empirical effectiveness of such an approach in terms of the treatment effect estimator over fixed effects models or more complex mixed effects models has not been examined. Further such research has been limited to the dichotomous treatments and has not considered continuous treatments. Moreover, such research has offered little concerning the inclusion of covariates with various relationships in specifying multilevel PS's and what role they may play in multilevel PS's. In particular, common approaches such as including every available variable often become intractable when estimating complex mixed effects models as the number of cross level interactions and random effects grow quickly. Even when including all the available variables in the PS represents a feasible approach, estimator variance with finite sample sizes can play a significant role in estimating treatment effects. Literature has shown that including effects related to the treatment but not to the outcome can increase the variance of the estimator considerably (Brookhart et al., 2006). As multilevel PS models may consider treatment assignment, in part, via random effects that may be related in various degrees to the outcome, it is important to understand how treatment assignment random effects relate to the outcome random effects and thus affect estimation of the treatment effect.

Research Questions

In this study I developed general framework for inferring multilevel treatment assignment mechanisms and assessed the performance of several fixed effects and mixed effects PS models that attempt to parametrically estimate such multilevel mechanisms. In particular, I evaluated the performance of different structural and stochastic PS specifications in the estimation of a treatment effect in which the true treatment assignment mechanism is multilevel in nature. I focused such inquiry on the situations where the true treatment assignment mechanism is influenced by individual level covariates, group level covariates and random effects based on both the group nesting structure and their interactions with individual characteristics. As such, I allowed the influence of individual characteristics on treatment status to vary by group. Using this treatment assignment mechanism and a nested outcome, I examined four questions:

1. How can we differentiate the different multilevel treatment assignment mechanisms and their corresponding propensity scores models?
2. How can we extend the multilevel propensity score framework to continuous treatments?

3. To what extent does the structure (e.g. fixed, mixed effects) of the propensity score model influence the properties of the treatment effect estimator when treatment assignment is multilevel in nature?
4. To what extent does variable specification of the multilevel propensity score affect estimates of the treatment effect?

In addressing the first two questions I considered five different PS model structures that may be considered to account for influence of the group on the individuals treatment assignment.

1. Single level logistic PS model that considers individual and group characteristics but does not account for nested group effects (Single level).
2. Single level logistic PS model that considers individual and group characteristics and accounts for nested group effects through fixed effects (Single level with fixed group effects).
3. Multilevel PS model that utilizes random group effects for the intercept only in addition to both individual and group covariates (Simple multilevel).
4. Multilevel PS that utilizes random group effects for the intercept only but allows non-randomly varying slopes (cross level interactions) in addition to both individual and group covariates (Moderate multilevel).
5. Multilevel PS model that utilizes random group effects for intercepts and slopes in addition to both individual and group covariates (Complex multilevel).

I addressed the second set of questions by considering three different types of covariates at each level. Allow *X* to indicate individual covariates and *W* to indicate group level covariates. The first type of covariate (X_1, W_1) represents true confounders in that it is both related to the outcome of interest as well as the treatment assignment. The second type (X_2, W_2) represents those covariates related to the outcome but unrelated to the treatment assignment whereas the third type (X_3, W_3) represents those covariates unrelated to the outcome but related to the treatment. Using such covariates I examined the effect of including seven different covariate combinations in the PS specifications on the bias, variance and mean-squared error (MSE) of the treatment effect estimator:

1. Only variables related to the both the treatment and outcome (1).
2. Only variables related to the outcome (2).
3. Only variables related to the treatment (3).
4. Variables related to the treatment and outcome plus those related to only the outcome (4).
5. Variables related to the treatment and outcome plus those related to only the treatment (5).
6. Variables related to only the treatment plus those related only to the outcome (6).
7. All available variables (7).

Theoretical Framework

Whereas statistical inference aims at determining significant associations, causal inference investigates a cause and effect relationship and requires more elaboration of the mechanism(s) by which a treatment affects an outcome. Although statistical inference does not necessarily imply the causality of a mechanism, valid statistical inference is generally considered supportive evidence of a cause and effect relationship. Several causal models have been posited. However, the current dominant theory of causality utilizes Rubin's Causal Model (Rubin, 1974; Holland, 1986). Based on the concept of the counterfactual, two structures form the basis for this model: the theory of potential outcomes and the concept of a treatment assignment mechanism.

In the former, Rubin's framework suggests that every experimental unit has a unique potential outcome for each potential assigned group membership (e.g. one outcome for treatment and one for control). Based on these potential outcomes, the causal effect of the treatment is estimated by contrasting these different outcomes. More formally Rubin defines a causal effect as:

> … the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1 (Rubin, 1974, p. 689).

Though this simple model of causality serves as a powerful tool, it is hindered by the fundamental problem of causal inference (Holland, 1986). As the Rubin Causal Model (RCM) estimates the causal effect of a treatment on a single subject by contrasting the outcome when the subject was exposed to the treatment and the outcome without the treatment the fundamental problem surfaces as we can not observe outcomes to multiple treatments on the same subject over the same period of time (Rubin, 1974; Holland & Rubin, 1983; Rosenbaum, 1984).

In the case of two alternative treatment conditions, I use Z as the treatment indicator such that $Z = 1$ if the treatment of interest was received and $Z = 0$ otherwise. The causal effect, δ_i, on an outcome, Y, for the i^{th} experimental unit, is the difference between the potential outcomes. More explicitly the causal effect is the difference between the potential outcome when receiving the treatment $Y_i^{(1)}$ given $Z = 1$ and the potential outcome when receiving the control $Y_i^{(0)}$ given $Z = 0$,

$$\delta_i = Y_i^{(1)} - Y_i^{(0)} \tag{2.1}$$

where the superscripts denote the potential outcomes with respect to the treatment indicator. The fundamental problem makes observing individual causal effects impossible, as we can not observe a single person's response, for example, to both taking a drug and not taking a drug during the same time period. Although, this feature prevents us from observing the treatment effect on an individual, it does not preclude us from estimating the average causal effect of the treatment over a population. Though we can not estimate a unit specific causal effect, we can estimate the average causal effect, $E[\delta]$, using sample statistics to estimate the average potential outcome for each group. In particular, by randomly sampling experimental units from a population and randomly assigning them to treatment conditions, we can ensure each unit has an equal probability of being assigned to each treatment. As a result, contrasting the average group treatment effects produces an unbiased estimate of the average treatment effect. In general, if we let $Y^{(1)}$ and $Y^{(0)}$ be random variables that denote the potential outcomes associated with the treatments, the average causal effect is equivalent to the difference between the average potential outcome of exposing all units in the population to one treatment and that of exposing all units to the alternative treatment, $E[Y^{(1)}] – E[Y^{(0)}]$ where $E[\cdot]$ denotes expectation. First, subject to assumptions subsequently discussed, we can write

$$\begin{aligned} E[\delta] &= E[Y^{(1)} - Y^{(0)}] \\ &= E[Y^{(1)}] - E[Y^{(0)}] \\ &= E[Y^{(1)} \mid Z = 1] - E[Y^{(0)} \mid Z = 0] \\ &= E[\overline{Y}^{(1)} - \overline{Y}^{(0)}] \end{aligned} \tag{2.2}$$

(e.g. Winship & Morgan, 1999). However, obtaining such equalities and resulting unbiased estimates of average causal effects from a random sample is dependent on the tenability of two major assumptions.

The first assumption is the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958). This assumption requires “the observation [or potential outcome] on one unit should be unaffected by the particular assignment of treatments to the other units” (Cox, 1958). Furthermore, it assumes that the potential outcomes of an experimental unit are independent of the treatments assigned to other experimental units, and that there is a single version of each treatment (Rubin, 1986). In essence, this assumption specifies that each subject has exactly one potential outcome under each group assignment. For example, if subjects have multiple potential outcomes based on not only his or her *own* treatment assignment but also on the treatment assignment of others, then there is no longer a one dimensional causal estimand. In such a case a researcher might relax SUTVA by defining multiple causal estimands and appropriately contrast outcomes (e.g. Verbitsky & Raudenbush, 2004). This assumption in conjunction with random sampling ensures that

$$E[Y^{(i)} \mid Z = 1] = E[\overline{Y}^{(i)}] \tag{2.3}$$

in (2.2) for *i* in {0,1} when there are two treatments.

The second assumption is strong ignorability of treatment assignment. This assumption corresponds with the second structure in Rubin’s causal model (the concept of a treatment assignment mechanism). The treatment assignment mechanism is the process by which some subjects were assigned to the treatment group and others were assigned to a control condition. In experiments this assignment mechanism is known and is completely random. However, in observational studies the assignment mechanism has non-random components and must be appropriately considered to provide unbiased estimates of each of the potential outcomes. Appropriate control for the non-random assignment mechanism is known as strong ignorability. More precisely, the assignment mechanism is strongly ignorable if treatment assignment is independent of potential outcomes given measured pretreatment characteristics. One assumes that after taking into account measured pretreatment characteristics, the treatment assignment was random. In this way, Rubin's causal model attempts to approximate the experimental quality of balanced groups. Though SUTVA is often a realistic assumption in various observational study designs, the strong ignorability assumption is a directly untestable assumption that potentially threatens all causal inferences. Strong ignorability is a critical and large assumption that amounts to accepting the conditional independence of the treatment assignment and potential outcomes. This assumption ensures that all units (or stratified units) have a nonzero probability of being assigned to a treatment and this probability is constant for all experimental units in the population. As a result, we can write

$$E[Y^{(i)}] = E[Y^{(i)} \mid Z = 1] = E[Y^{(i)} \mid Z = 0] \tag{2.4}$$

in (2.2) for *i* in {0,1} when there are two treatments. Thus, given the validity of these assumptions, an unbiased estimate of the population average treatment effect can be obtained from the differences in sample means.

Though such an approach provides a statistical method for causal inference its validity and unbiasedness rests on the ignorability of treatment assignment. In particular, substantive theory or empirical evidence might suggest that the outcome and treatment assignment vary within a population based on pretreatment characteristics. In other words, the probability of being assigned to each treatment varies across the experimental units. As a result, the strong ignorability of treatment assignment assumption is violated. However, if those pretreatment characteristics that influence the treatment assignment are observed, one can restrict comparisons to those units that have similar probabilities of receiving treatment conditional on the pretreatment characteristics. Accordingly, we swap the assumption of unconditional ignorability of treatment assignment for the conditional ignorability of treatment assignment assumption. We now assume that the treatment one receives is conditionally independent of the potential outcomes $Y^{(1)}$ and $Y^{(0)}$ when the influential pretreatment characteristics are fixed. Under SUTVA and a random sample it can be shown that

$$E[\bar{Y}^{(1)} \mid \mathbf{X} = \mathbf{x}] = E[Y^{(1)} \mid Z = 1, \mathbf{X} = \mathbf{x}] = E[Y^{(1)} \mid Z = 0, \mathbf{X} = \mathbf{x}] = E[Y^{(1)} \mid \mathbf{X} = \mathbf{x}] \quad (2.5)$$

$$E[\bar{Y}^{(0)} \mid \mathbf{X} = \mathbf{x}] = E[Y^{(0)} \mid Z = 0, \mathbf{X} = \mathbf{x}] = E[Y^{(0)} \mid Z = 1, \mathbf{X} = \mathbf{x}] = E[Y^{0)} \mid \mathbf{X} = \mathbf{x}] \quad (2.6)$$

where $\boldsymbol{X}$ denotes the set of pretreatment covariates that confounds the causal effects of treatments on the outcomes. Departing from the unconditional approach, we estimate the average causal effect, δ, with sample statistics by first estimating the effects for each subpopulation defined by $\boldsymbol{X}$**.** Using (2.5) and (2.6)

$$
\begin{aligned}
E[\delta \mid \mathbf{X}=\mathbf{x}] &= E[Y^{(1)} - Y^{(0)} \mid \mathbf{X}=\mathbf{x}] \\
&= E[Y^{(1)} \mid \mathbf{X}=\mathbf{x}] - E[Y^{(0)} \mid \mathbf{X}=\mathbf{x}] \\
&= E[Y^{(1)} \mid Z=1, \mathbf{X}=\mathbf{x}] - E[Y^{(0)} \mid Z=0, \mathbf{X}=\mathbf{x}] \\
&= E[\bar{Y}^{(1)} \mid \mathbf{X}=\mathbf{x}] - E[\bar{Y}^{(0)} \mid \mathbf{X}=\mathbf{x}]
\end{aligned}
\tag{2.7}
$$

Second, assuming approximate homogeneity of the treatment effect, we average the causal effect estimates across the groups defined by ***X*** weighting by the density of ***X***. Through the RCM we can make explicit the assumptions needed for causal inference. Such a framework helps understand and identify those causal questions that are tractable. Furthermore, this framework provides a strong conceptual basis from which we can explore statistical methods to infer causality and estimate causal effects.

Of particular interest in this dissertation are those statistical methods within the RCM framework that are used to study the effects of treatments in observational data. In observational data a researcher makes no attempt to manipulate the situation generating the data. As a result, the fundamental problem with observational data is that, for example, students and schools choose their situations or treatments according to some criteria. Accordingly, in estimating causal effects, researchers must adjust for all factors that led the individual or school to their choice that might also be correlated with the outcome of interest. To appropriately adjust for the factors influencing choice or treatment assignment, researchers have developed a variety of methods to support causal inferences using observational data.

Traditionally, researchers have used ordinary least squares regression (OLS) to study the impact of school resources on outcomes (e.g., Coleman et al 1966). In the cross-sectional case, the analyst relies on a specification such as,

$$Y = X\beta + Z\delta + \varepsilon \tag{2.8}$$

where $\boldsymbol{X}$ represent the control variables the analyst is attempting adjust for, $\boldsymbol{\beta}$ are the corresponding coefficients of the control variables, Z is the treatment assignment with coefficient δ, and ε is the error. However, among other assumptions, such an approach assumes, that all confounding variables have been measured and that there is a specific parametric relationship between the treatment and the outcome to which OLS is sensitive to. Further, such assumptions rely heavily on extrapolation in that they frequently estimate the counterfactual by extending the regression line beyond the scope of the data. As a result, OLS estimates are likely inaccurate in a variety of settings.

In a similar approach, regression discontinuity relies on the existence of a treatment assignment rule or cutoff. In particular, in regression discontinuity designs participants are assigned to a treatment groups based solely on a particular pretreatment characteristic. Those participants that are above the cutoff receive one treatment and those below the cutoff receive another. Such designs often supply inferences similar to randomized experiments in that the treatment assignment is known to be unrelated to all confounders except that which determined the treatment. Though such a design is formidable, in observational data, rarely does a single pre-treatment variable decide the treatment assignment and rarely is there such a sharp or even fuzzy cutoff such that all above a certain value received the treatment.

Another possible method for causal inference in observational studies is an instrumental variable approach. Such an approach relies on the existence of an instrument or in other words a variable correlated with the treatment but uncorrelated with the outcome conditional on other covariates. This approach attempts to identify variables that would not be expected to independently alter outcomes but do so only through an endogenous measured variable. Though this approach offers unbiased estimation of causal effects when a number of assumptions are met, its utility on observational data is reliant on the existence of a high quality instrument. Consequently, its use is generally limited to situations in which there are identifiable and measured exogenous events, e.g. certain policy changes, which have no direct effect on the outcome but rather work exclusively through an endogenous variable.

A fourth possible method for causal inference in observational data is the PS and is the focus of this investigation. In particular, the focus of this study is causal inference in observational studies when a high quality instrumental variable or pretreatment cutoff variable is unavailable and PS based approaches provide a reasonable approach. PS based methods conceptually attempt to identify and contrast similar units to estimate a causal effect. Though such an approach has considerable flexibility in that it does not require a type of quasi-experimental setting that prior approaches did, it requires other assumptions. For instance, it assumes that one can reasonably infer the treatment assignment from the measured covariates and that all covariates that influenced the treatment assignment were measured.

Causal inference in observational studies with PS based approaches within the RCM conceptually attempts to mimic a randomized experiment in which the treatment assignment mechanism is known to be a function of measured pretreatment covariates $\boldsymbol{X}$. In assuming this approach with observational data, the primary task then is to infer the treatment assignment mechanism from the measured data. In the case were the treatment assignment can be reasonably inferred

In the case of two treatments, the PS represents the conditional probability of assigning each experimental unit to the treatment of interest. That is, assuming two different treatment conditions where Z=1 represents the treatment of interest and Z=0 represents some control condition, the PS, *e(**X**)*, is

$$e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X}) \tag{2.9}$$

Accordingly, if adjustment on all measured covariates is sufficient for unbiased estimation of the treatment effect, then so is adjustment on the PS (Rosenbaum & Rubin, 1983a). That is, treatment assignment is conditionally independent of potential outcomes give the PS. The PS acts as a unidimensional balancing score in which all the information relevant to balancing treatment assignment in ***X*** is extracted in *e(**X**)*. As a result, conditioning on *e(**X**)* balances the distributions of ***X*** between the treatment and control groups and thus ensures the strong ignorability of treatment assignment assumption needed for causal inference in the RCM. Accordingly, units with similar PS values but different treatment assignments can serve as counterfactuals estimates for the missing potential outcome. In other words, the expected difference in the observed responses to different treatment conditions when the PS is held fixed is an unbiased estimate of the average treatment effect.

In a similar manner, the case of a continuous single level treatment has been extended as the generalized propensity score (GPS) (e.g. Imai & Van Dyk, 2004). Whereas in the dichotomous treatment situation there are only two potential outcomes under the RCM, in the continuous treatment case there are a set of outcomes. More formally, the GPS framework first assumes that Z is continuously distributed measure,

$\{Y_i^{(z)}\}\, for\, z\, in\, Z$, Z and $\boldsymbol{X}$ are defined on a common probability space and that $Y = Y(Z)$ is a well defined random variable (Hirano & Imbens, 2004). Under such assumptions, for unit *i*'s outcome *Y* and treatment assignment *z,* we can define the unit-level dose response function as

$$Y_i^{(z)}\ for\ z \text{ in } Z \tag{2.10}$$

In evaluating the causal effect of a continuous treatment our interest lies in the average dose-response function

$$\mu(z) = E[Y_i^{(z)}] \tag{2.11}$$

Hirano & Imbens (2004) define the GPS as

$$E = e(Z, \vec{X}) \tag{2.12}$$

where

$$e(z, \vec{x}) = f_{Z|\vec{X}}(z \mid \vec{x}) \tag{2.13}$$

and $e(z, \vec{x})$ is the conditional density of the treatment given the covariates. In other words the GPS represents the conditional probability of assigning each experimental unit to one of the levels of the treatment of interest. Similar to the standard PS, the GPS has a balancing property such that within strata defined by the GPS, the probability that *Z*=*z* does not depend on the value of the covariates ***X***. In other words when matching exactly on the GPS

$$\vec{X} \perp I\{Z = z\} \mid e(z, \vec{X}) \tag{2.14}$$

where *I* is an indicator. As a result, the conditional distribution of the treatment assignment is independent of the outcomes

$$f_Z(z \mid e_i(z, \vec{X}), Y_i^{(z)}) = f_Z(z \mid e_i(z, \vec{X})) \tag{2.15}$$

That is, if ignorability of the treatment assignment with respect to the potential outcomes given the measured covariates is valid, then the potential outcomes will be independent of the treatment assignment given the. Consequently the GPS is similar to the PS in that it helps satisfy the RCM by providing strong ignorability of treatment assignment.

In observational studies, though $\boldsymbol{X}$ have been observed, the PS is generally unknown and needs to be estimated from the data. Though we generally do not have the true PS in observational data, estimated PS's tend to operate like true PS's in that comparing units on a fixed score tends to balance covariate distributions between treatment groups (Rosenbaum & Rubin, 1983a). In particular, theory has suggested that use of the empirical PS is often more effective than use of the actual PS as it tends to remove empirical confounders or chance imbalances due to sampling variability (Robins, Mark & Newey, 1992; Rosenbaum, 1987).

Uses of Propensity Scores

The literature surrounding the use of PS's has proposed several alternative uses of PS's for causal inference. In particular, since the PS is a tool to identify comparable units by balancing the distribution of their pretreatment covariates, the PS needs to be used to contrast treatment levels. In this project, I consider three alternative uses of PS's: subclassification, matching and inverse probability of treatment weighting (IPTW).

Stratification on the Propensity Score

A first use of the PS is to separate the experimental units in to subclasses of similar PS values. Such division creates similar covariate distributions among the treatment groups within a subclass. As a result, within a subclass the observed responses of the control units provide a reasonable basis for inferring the counterfactual responses of the treatment units. Typically, subclassifying on fives classes of equal size removes approximately 90% of the bias associated with the measured covariates Rosenbaum & Rubin, 1983a; Cochran, 1968).

Matching on the Propensity Score

A second use of the PS is to match experimental units on the basis of the PS. The intention with this use is to create comparable sets of treated and control subjects. Similar to stratification, matching units on the PS creates similar covariate distributions among the treatment groups. As a result, the matched control unit responses provide a reasonable basis for inference on the counterfactual responses for the matched treatment units. Though exact matching on the PS is optimal, approaches such as matching the nearest available neighbor are utilized to make inference tractable (e.g. Rosenbaum, 1989). In particular, this use is most appropriate when there is a large reservoir of potential control units available. Though greedy matching schemes such as nearest neighbor provide simple approaches, it may not be optimal in terms of minimizing differences within matches (Rosenbaum, 1993). An alternative algorithm which attempts to minimize such global differences within matches is full matching (Hansen, 2004; Rosenbaum, 1991). In particular, this algorithm uses network flow designs (Hansen & Klopfer, 2006) and results in the smallest average distance within matched sets and contains one or more subjects from each treatment group in each matched set. In this study, I focused on the use of full matching as implemented in the R package *optmatch* (Hansen & Klopfer, 2006) to study multilevel PS's used for matching in HLMs.

Weighting on the Propensity Score

A third use of the PS is to weight by the inverse probability of receiving the treatment (Robins, Hernan & Brumback, 2000). In particular, weights are constructed for experimental units by first estimating their probability of receiving treatment and then weighting them by the inverse of the probability in a parametric or non-parametric procedure. Such an approach creates a pseudo-population for each treatment group through weighting by the inverse probability of receiving the treatment. Under strong ignorability, the weighted mean difference between the treatment groups is a consistent estimate of the average treatment effect. However, such an approach relies heavily on the estimated weights and is easily influenced by the estimation and parametric structure of both the propensity model and the outcome model when used.

Combining PS and Parametric Outcome Models

The PS is a method designed to provide ignorability of the treatment assignment rather than directly estimate a treatment effect. In adopting one of the three above PS uses researchers subsequently evaluate the average treatment effect by contrasting the appropriate outcomes. In doing this one may additionally utilize a parametric or non-parametric structure to model the conditional relationship between the treatment and the outcome (e.g. Rosenbaum & Rubin, 1983a). When parametric structures are appropriate, research has demonstrated benefits from combining the PS with, for example, regression adjustment (e.g. Hirano & Imbens, 2002; Kleyman & Hansen, 2008). Moreover, Robins and Rotnizky (1995; Robins, Rotnizky & Zhao, 1995) demonstrated that as long as only one of the models, either that for the conditional mean of the potential outcomes given covariates, or that for the treatment variable given the covariates, is correctly specified, the resulting estimator will be consistent. Of particular interest to this study and educational research in general, is addressing the multilevel nature of many educational phenomena. In particular, because students within the same classroom share the same teacher and school, we would like our treatment effect estimator to take into account the lack of independence between students. Consequently, this study focuses on adjusting for imbalances through the three different PS uses above within the context of a parametric multilevel model (e.g. Correnti & Rowan, 2007). Such an approach combines the estimated PS with a standard parametric HLM to address the nonrandom treatment assignment and the multilevel nature of education using a linear approximation of achievement.

Observational Study Design & Model

In educational research and other fields, research data often have a hierarchical structure. That is, the individual subjects of study may be classified or arranged in groups which themselves have qualities that influence the study. In this case, the individuals can be seen as the first level of units in the study and the groups into which they are arranged are second level units. Indicated by the questions and focus of this study, I concentrated on estimating PS for use in multilevel outcome models. To address the nested structure of the outcome, I utilized a hierarchical linear model (HLM) (Raudenbush & Bryk, 2002). In HLMs each level of the nested structure is formally represented by its own sub-model. For example, in a two level model where students are considered to be nested within schools, we can represent the level one student model as

$$Y_{ij} = \pi_0 + \sum_{p=1}^{P} \pi_p X_{ij} + \varepsilon_{ij} \tag{2.16}$$

where Y represents an outcome such as math achievement, π_0 is the average student score adjusted for the student variables, X, and the corresponding coefficients, π_p while ε has a normal distribution with mean zero and variance σ^2. To link the students and schools, we can represent the school through a sub-model or level two school model as

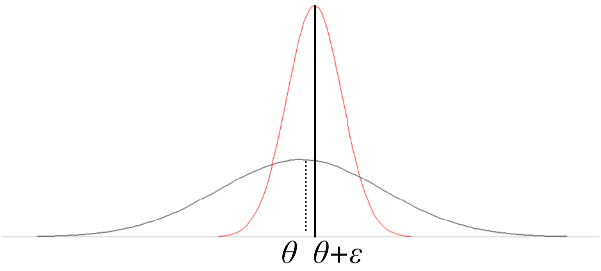
$$\pi_0 = \beta_{00} + \sum_{q=1}^{Q} \beta_{0q} W_{qj} + r_{0j} \tag{2.17}$$

where β_{00} is the average adjusted achievement for school, β_{0q} is average effect of covariate, W_q, on adjusted achievement with corresponding coefficients, β_{0q} and r_{0j} is the random effect of school j and has a normal distribution with mean zero and variance τ_π. These sub-models articulate relationships among covariates within a given level and, in turn, express how variables at one level influence relations occurring at other levels (Raudenbush & Bryk, 2002).

As PS's intend to mimic randomized experiments, I focus this study on the HLM multi-site observational designs most closely resembling multi-site randomized designs. In multi-site randomized designs the experimental units are individuals within each site. In particular, within each site individuals are assigned to different treatments. Though such assignment is at random in an experiment, in an observational study both the individual and group potentially contribute, in various ways, to each individual's treatment assignment. Consequently, if both the group and individual influence the treatment assignment, the correct PS will be a function of both individual and group covariates.

Variable Selection in Propensity Scores

A primary utility of the PS model approach is its potential ability to mimic randomization by making the treatment assignment conditionally independent of potential outcomes given the observed covariates. However, this utility is often tempered by the difficulty of correctly specifying the PS. In particular, the bias and variance of the treatment effect estimator strongly depend on the subset of observed variables included in the construction of the PS, especially in finite sample sizes. Consistent estimators ensure that the variance and bias of the treatment effect estimator goes to zero as the sample size approaches infinity. However, such consistent estimators in finite sample studies provide much less protection from the variance of an estimator as chance imbalances are much more likely. Relevant to education studies with hierarchical outcomes and group level treatments, such variability of the treatment effect estimator often plays a large role as the effective sample size depends on the number of groups rather than individuals. Consider Figure (2.18), in which, for a given sample size, two different consistent estimators are graphically compared. Though the first estimator, θ_1, is unbiased, its density is thoroughly dispersed throughout the parameter space indicating the estimator's variability. In contrast, the second estimator, θ_2, is slightly biased but its density is concentrated around its center. Consequently, we are forced to develop a criterion that accounts for both bias and variance in order to evaluate which estimator is more appropriate.



Figure(2.18): Density of two different estimators: Black is unbiased but fairly dispersed and red is slightly biased but concentrated around the estimand, θ

To attend to this tradeoff, Rubin and Thomas (1996) derived approximations for the reduction in the bias and variance of an estimated treatment effect using the PS. Such derivations support including all variables related to the outcome regardless of their relationship with the treatment assignment. Additionally, their derivations demonstrated that including variables that are strongly related to treatment assignment but unrelated to the outcome can increase the variance of the estimator without a corresponding decrease in bias. Accordingly, theoretical literature has suggested: (1) including variables unrelated to the treatment assignment but related to the outcome and (2) the exclusion of variables that are related to the treatment assignment but unrelated to the outcome as such an approach decreases bias without increasing variance (Rubin & Thomas, 1996; Brookhart et al., 2006). In other words, one should exclude those variables resembling the properties of an instrumental variable and if such a variable can be conceptualized as a high quality instrument consider an instrumental variable approach. To attend to the tradeoff between bias reduction and variance inflation caused by variable selection in the PS, such literature has considered the mean-squared error (MSE) of the treatment effect estimators corresponding to various variable choices (e.g. Brookhart et al., 2006). In particular because MSE summarizes the contribution of both the bias and variance of an estimator it can represent the quality of an estimator. That is,

$$MSE = (\text{bias}(\hat{\theta}))^2 + \text{var}(\hat{\theta}) \tag{2.19}$$

Despite literature assessing the structure and specification of such single level outcome and PS models, little is known concerning the performance of such strategies in multilevel situations with finite sample sizes.

Though there are practical strategies to constructing the PS such as a stepwise approach, constructing the PS model in such ways may impair the ability of the PS to contrast meaningfully comparable groups as they exclusively focus on the treatment without consideration for the outcome. Consequently, they neglect the duality of confounding by ignoring the effects of variables that are related to the outcome but weakly related to treatment assignment. Such neglect often results in increased treatment effect estimator variance without a corresponding decrease in bias. Similarly, although including a variable that has little to no relationship with the outcome but a strong relationship with the treatment does not bias the estimator, it can add substantial variance to the estimator. Though most studies use treatment effect estimators that are consistent, adding such variance in finite sample sized studies can detract significantly from the quality of the estimator. To better understand the bias–variance tradeoff in constructing multilevel PS's I assessed the performance of PS's with different covariate combinations in estimating the treatment.

Methods

Developing and Differentiating Propensity Scores for Multilevel Mechanisms

I first attend to the how we might conceptualize the involvement of the group in the selection processes. Next, I pose several models to estimate such mechanisms and discuss their implications and additional considerations. Third I extend the multilevel PS framework to include continuous treatments. Next, I discuss its implications and additional considerations including how the score may be used. Finally, I assess the performance of the PS's in conjunction with different sets of variables.

In single level PS analyses researchers assume that the potential outcomes are independent of the treatment assignment given the measured covariates. That is,

$$Y^{(i)} \perp Z \mid \vec{X} \qquad (2.20)$$

Using the properties of the PS, one can achieve similar strong ignorability via the PS in that the potential outcomes will also be independent of the treatment assignment given the PS

$$Y^{(i)} \perp Z \mid e(\vec{X}) \tag{2.21}$$

However, in studies where the treatment assignment mechanism involves some influence from the group an individual belongs to, (2.21) may no longer hold. As a result we must now take into account the role of the group in deciding treatment status

$$Y^{(i)} \perp Z \mid \vec{X}, U_X, \vec{W}, \vec{U}_W, \vec{r} \tag{2.22}$$

where $\vec{X}$ represent the observed individual level covariates, $\vec{U}_X$ represent the unobserved individual level covariates, $\vec{W}$ represent the observed group level covariates which include group membership, $\vec{U}_W$ represent the unobserved group level covariates and $\vec{r}$ represent random effects of the group. Since (2.21) is no longer a reasonable basis for assuming ignorability, we must now specify the PS or the probability of receiving the treatment as a function of both individual and group membership and characteristics. Whereas

$$P(Z = \mathrm{z}) = f(\vec{X}, \vec{U}_X) \tag{2.23}$$

estimated the PS in the single level case, we now must rely on

$$P(Z = z) = f(\vec{X}, \vec{U}_X, \vec{W}, \vec{U}_W, \vec{r}) \tag{2.24}$$

to specify the PS and account for the group level influence.

Tantamount to estimating multilevel treatment assignment mechanisms is accepting that the influence of pretreatment covariates on the treatment assignment may differ substantially between groups and/or individuals. In applying a multilevel approach to the PS, we must consider the context and processes of the multilevel settings as the manner in which differences in selection processes manifest may dictate the approach needed (Kim & Seltzer, 2007). The main consideration is how the group selection processes actually differ. Specifically, we need to identify whether, simply, the average treatment level varies between groups or, more complexly, if the average treatment level varies and the magnitude of slopes linking individual covariates to treatment vary (Kim & Seltzer, 2007). For instance, in hierarchical linear models, these differences are often referred to as a random intercept model and a random intercept and random slope model, respectively. Failure to account for such differences in the selection mechanisms, when they exist, will bias the treatment estimator. In addition, though there may be certain loss of efficiency in accounting for such differences when they do not exist, this does not bias the estimator. Furthermore, as subsequently affirmed in other contexts, the potential loss of efficiency from utilizing a multilevel structure when it is untrue is most likely dominated by the potential bias added by ignoring the multilevel structure.

Differentiating Mechanisms

In the context of education, attending to how treatment selection mechanisms differ among groups, estimating a common single level PS and restricting comparisons to individuals within the same group has been a historical approach (Rosenbaum, 1986). Such an approach attempts accounts for group effects by only contrasting individuals within the same group. Restricting comparisons to individuals within group on any multilevel PS presents an attractive approach as it has several utilities. First, restricting comparisons to within groups preserves the natural hierarchical relationships and allows us to contrast comparable units. For example, depending on the true selection mechanism, matching within groups may potentially adjust for unobserved group level covariates as they are common within matches. Another advantage of restricting comparisons to within groups is it facilitates the study of variation in treatment effect between groups (Hong & Raudenbush, 2006; Kim & Seltzer, 2007). In particular, if our focal research question is directly interested in the extent to which the treatment effect varies between groups, then it is useful, and perhaps more accurate, to consider within group matches so as to explicitly balance covariates within each group and ensure more accurate estimates of the treatment effect at each group.

The equivalence of such an approach and a multilevel PS that explicitly considers group membership, and thus variation in the selection mechanism, is dependent on the true selection mechanism. In particular, if the selection mechanism acts such that group membership and covariates contribute solely to the average treatment level probabilities, then restricting comparisons to within groups based on the single level PS that ignores group membership is an effective way to control for group selection bias (Kim & Seltzer, 2007). This approach blocks on group membership, and thus on observed and unobserved group covariates, and constrains interactions between group covariates and membership with individual covariates to be zero. As a result all contributions from the group membership cancel out within comparable sets and this will be an effective approach in providing unbiased and consistent estimates of the treatment effect. Although this approach potentially addresses the group's role in treatment assignment in certain circumstances, it has several practical and theoretical limitations. Practically, the utility of within group comparisons is often mitigated by the lack of comparable individuals within each group. In particular, restricting comparisons to within group memberships generally requires a large reservoir of treatment and control individuals in each group. Studies that examined the effects of treatments that are uncommon or are enacted at the individual level often constrain the pool of potential comparisons as it can be difficult to identify comparable individuals within groups. Theoretically, using a common single level PS, even if restricting comparisons to within groups, assumes that the selection processes are fixed among groups up to a constant for each group. In other words, in a parametric model, a common single level PS would assume that the coefficients do not vary, whether in a fixed or random manner, among groups but rather only the intercept does. However, under other circumstances mechanisms may allow the effects of individual covariates differ among groups as each group weights the cross level interactions differently and such cross level interactions are modified by group level covariates thereby producing different scores. In general this will result in different comparisons though Kim and Seltzer (2007) note the unique case when only one slope is random and certain comparisons such as nearest neighbor matching are used. Under this unique condition matching within groups using a single level PS will produce the same matches as those multilevel PS's that address the varying coefficients and cross level interactions as the numerical valued propensities will differ in value but absolute ranking will remain consistent. Consequently, if the influence of individual characteristics on the treatment assignment vary among groups or are modified by the group characteristics, ignoring the clustering and/or cross level interactions between covariates and group memberships, even when using within group comparisons, can misidentify comparable units and provide spurious inferences.

To address the practical and theoretical limitations of the single level approach, literature has proposed several parametric options focusing on hierarchical generalized linear models (HGLMs) (Hong & Raudenbush, 2006; Kim & Seltzer, 2007; Raudenbush & Bryk, 2002). Such approaches tend to allow for more flexibility in that they allow the selection processes to differ between groups. In developing a framework for multilevel PS's I considered five different ways in which the selection processes may differ based on past literature and posed an additional process. For each of these processes I considered a different parametric model. The first model is a common single level model that incorporates both the individual covariates as well as the group level covariates but does so by ignoring group membership. Further, by ignoring group membership the single level assumes individuals in the same group to be independent of each other. In the dichotomous treatment case, this can be estimated using the logistic regression model within the generalized linear model framework. Using matrix notation we can specify this model as

$$log(\frac{p}{1-p}) = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} \tag{2.25}$$

(McCullagh & Nelder, 1983). In particular, in using the single level model, one assumes that the true treatment assignment mechanism is uninfluenced by group membership and that group characteristics contribute to an individual's propensity in a common manner across all groups. That is, an individual's group membership offers no information concerning his or her treatment assignment and further such group membership does not modify the influence of any individual characteristics as they are also constant across all groups.

The second model type I considered was a single level fixed effects regression model. Specifically, this approach advances the single level model only in that it includes a series of indicator variables for group membership in conjunction with the individual and group characteristics

$$log(\frac{p}{1-p}) = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{I}\vec{\pi} + \vec{\varepsilon} \qquad (2.26)$$

where $\vec{I}$ is an indicator matrix corresponding to group membership. As this model is saturated with respect to the group memberships, it does tend to account for group influence. However, this model still neglects the potential differential functioning of individual covariates in influencing the treatment assignment. The third model type I considered was a HGLM or a generalized linear mixed effects model with the previous fixed effects and only a random intercept effect (e.g. Pinheiro & Bates, 2000). This approach considers the influence of the individual and the group as well as the group membership and the dependence of individuals within a group. In mixed form it is

$$log(\frac{p}{1-p}) = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{r}_0 \qquad (2.27)$$

where $\vec{r}$ is a single random effect for each group such that $\vec{r}_0 \sim N(0, \vec{\tau})$. Under this particular model, one assumes that the group influences the treatment assignment of its members in a common way. In a two level example where students are nested in school and treatments are assigned to students under this mechanism, the school has a constant and uniform influence on each of its students. In a practical example where we consider individual tutoring the treatment, this model might be appropriate if school membership and characteristics increase or decrease each of its students chances of receiving tutoring in a common way.

Fourth I consider another HGLM or generalized linear mixed effects model with the previous fixed effects and single random effect for the intercept but also cross-level interactions between the covariates. In mixed notation we have

$$log(\frac{p}{1-p}) = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{X}\vec{W}\vec{\pi} + \vec{r}_0 \tag{2.28}$$

This model permits more flexibility as to the assumptions of the groups' influence. In particular, this model assumes that the group has a shared common influence on all of its individuals as before but also that the group characteristics interacts with individual characteristics to modify the role of such individual characteristics in predicting treatment assignment. For example, if English tutoring was the treatment, schools with a higher percentage of teacher certified in teaching English as a second language may increase the odds that students in their schools with limited English proficiency receive the treatment as these teacher may an increased perceived benefit for these students. In other words, groups may now additionally influence certain subsets of their individuals by inflating or deflating their probabilities of receiving treatment. The fifth model was a generalized linear mixed effects model with previous fixed effects and cross-level interactions between the covariates and now a random group effect for the each individual covariate. In mixed model form we can specify this as

$$log(\frac{p}{1-p}) = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{X}\vec{W}\vec{\pi} + \vec{X}\vec{r} \tag{2.29}$$

where $\vec{r} \sim MVN(0, \vec{\tau})$. In other words, this model considers the treatment assignment mechanism to be composed of individual influence (*X*), a common group influence (*W, r_0*), a differential influence on subgroups of individuals that is common across all groups

(*XW*) and now an additional differential influence on subgroups that is unique to each group (*Xr*).

The multilevel models posed represent an increasing complex treatment assignment mechanism. Accordingly, I term then simple, moderate and complex. That is a simple multilevel treatment assignment mechanism can be modeled using (2.27); a moderate multilevel treatment assignment mechanism can be modeled using (2.28); and a complex multilevel treatment assignment mechanism can be modeled using (2.29). Further, as the first two models posed are not strictly speaking multilevel models I call them a common single level model (2.25) and a common single level model with fixed group effects (2.26).

Each mechanism engages variation at the group level, however, they maintain different assumptions. The common single level assumes that group membership is does not contribute any variation in treatment assignment but group level covariates do. The common single level with fixed group effects assumes that both the group membership and group covariates contribute and that observations are independent. Further each multilevel mechanism engages variation meaningfully different ways. For instance, simple multilevel PS assumes the contributions of the covariates from the individual level to the probability of treatment assignment as fixed across all groups. In such a case, the variation in the selection process is solely a function of group membership. Increasing the complexity of the mechanism, the moderate PS assumes variation comes from group memberships and the interaction between group covariates and individual covariates. Further, the complex PS assumes variation comes from group memberships, the interaction between group covariates and individual covariates and additionally the interaction between group membership and individual covariates. The difference between the processes can be conceptualized as the simple mechanism implying that all members of the same group share the same shift in probability of receiving a treatment level regardless of their personal characteristics and that shift is determined by the group covariates only. In the moderate case, this mechanism implies that all members of the same group receive a shift in probability based on that common membership and an additional shift in probability based on the interactions between individual covariates and the group characteristics. Finally the more complex mechanism implies that all members of a the same group receive a shift in probability based on that common membership, an additional shift in probability based on the interactions between individual covariates and the group characteristics and yet another shift in probability based on the interactions between individual covariates and group membership. Under the presence of the complex mechanism, cross-level interactions between group membership and individual covariates influence the probability so that each group has a different PS equation.

In applying multilevel PS's, utilizing such models requires thoughtful consideration of how the units were selected into treatments. Theoretically assessing whether a selection process implies that only the average treatment level varies between groups or if the average treatment level varies and the magnitude of slopes vary can be difficult especially when dealing with treatments that are not well studied. However, in observational studies, because treatment assignment is not necessarily being directly studied, the essential purpose of constructing a PS remains to make the potential outcomes independent of the treatment assignment rather than to test hypotheses concerning the treatment assignment. As a result, empirically assessing the treatment assignment mechanism often aligns with the PS purpose. Because the multilevel PS's represent nested models in the sense that the smaller simpler models constrain certain complex, larger parameters to be zero, we can empirically assess which mechanism most closely aligns with the selection process. Further, the mechanism that most closely aligns with the observed selection process theoretically most closely mimics randomization in that the treatment assignment will be random within comparable sets individuals. As a result, the identified mechanism should theoretically provide the most bias reduction in the treatment effect estimator without adding unnecessary variance.

Next, to address the nature of various treatments in education, I developed a framework to utilize the PS when the treatment is continuous and is influenced by both individuals and groups. In particular, drawing on literature, I extended the dichotomous multilevel PS to continuous treatments by combining the GPS with the dichotomous multilevel PS (Hong & Raudenbush, 2006; Kim & Seltzer, 2007; Hirano & Imbens, 2004). That is, how might one construct the PS in multilevel settings when the treatment represents a continuous or dosage treatment?

Following Hirano & Imbens (2004), I first assume that *Z* is continuously distributed measure, $\{Y_i^{(z)}\}\, for\ z\, in\, Z$, *Z* and ***X*** are defined on a common probability space and that $Y = Y(Z)$ is a well defined random variable (Hirano & Imbens, 2004). Under such assumptions, for unit *i*'s outcome *Y* and treatment assignment *z,* we can define the unit-level dose response function as

$$Y_i^{(z)}\ for\ z \text{ in } Z \tag{2.30}$$

Further, I assume our interest in evaluating the causal effect of a continuous treatment lies in the average dose-response function

$$\mu(z) = E[Y_i^{(z)}] \tag{2.31}$$

I then define the multilevel GPS for continuous treatments as

$$E = e(Z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r}) \tag{2.32}$$

where

$$e(z, \vec{x}, \vec{w}, \mathrm{x}\vec{w}, \vec{x}\vec{r}) = f_{Z|\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(z \,|\, \vec{x}, \vec{w}, \mathrm{x}\vec{w}, \vec{x}\vec{r}) \tag{2.33}$$

and $e(z, \vec{x}, \vec{w}, \mathrm{x}\vec{w}, \vec{x}\vec{r})$ is the conditional density of the treatment given the covariates. Further, as before, *z* is the treatment assignment, *x* are the individual level covariates, *w* are the group level covariates, *xw* are the cross level interactions and *r* are the random effects which account for group membership and interact with the individual covariates (*xr*). In other words the multilevel GPS represents the conditional probability of assigning each observational unit to one of the levels of the treatment of interest. Similar to the standard PS, the multilevel GPS has a balancing property such that within strata defined by the multilevel GPS, the probability that *Z*=*z* does not depend on the value of the covariates ***X***. In other words when matching exactly on the multilevel GPS

$$\vec{X} \perp I\{Z = z\} \,|\, e(z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r}) \tag{2.34}$$

where *I* is an indicator. Next assume that we have measured all variables at both the individual and group level that influence the treatment assignment such that the potential outcomes are independent of the treatment assignment given the pretreatment covariates. Formally,

$$Y_i^{(z)} \perp Z \,|\, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r} \tag{2.35}$$

for all z in Z. Similar to the canonical PS, it can be shown that the conditional distribution of the treatment assignment given the multilevel GPS is also independent of the potential outcomes (Appendix A). Formally,

$$f_Z(z \mid e_i(z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r}), Y_i^{(z)}) = f_Z(z \mid e_i(z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r})) \tag{2.36}$$

That is, if ignorability of the treatment assignment with respect to the potential outcomes given the measured covariates is valid, then the potential outcomes will be independent of the treatment assignment given the multilevel GPS. Consequently the multilevel GPS is similar to the canonical PS and GPS in that it helps satisfy the RCM by providing ignorability of treatment assignment.

In a manner similar to that of the multilevel PS for a dichotomous treatment, we can conceptualize the treatment assignment mechanism in at least five ways. However, as treatment now represents a continuous outcome, we can parametrically estimate the PS or the conditional density of the treatment assignment mechanism using models for continuous treatments. That is, a possible way to estimate the single level model (2.25) is

$$\vec{Z} = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{\varepsilon} \tag{2.37}$$

The single level fixed model (2.26) now becomes

$$\vec{Z} = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{I}\vec{\pi} + \vec{\varepsilon} \tag{2.38}$$

Further the multilevel simple (2.27), moderate (2.28) and complex models (2.29) can now be represented by

$$\vec{Z} = \vec{X}\vec{\beta} + \vec{W}\vec{\gamma} + \vec{X}\vec{W}\vec{\pi} + \vec{X}\vec{r} + \vec{\varepsilon} \tag{2.39}$$

where $\vec{\pi}$ is constrained to be zero in the simple case, and $\vec{r}$ is a single random effect for group membership in the simple and moderate cases but represents multiple random effects in the complex case.

Efficacy of Multilevel PS's & Role of Covariates in Multilevel PS's

There is theoretical and practical motivation for developing multilevel PS's when multilevel treatment assignment mechanisms exist. Theoretical motivation stems from the potential of the group to influence individual's treatment assignments and to do so in different ways. Practical motivation stems from limitations such as the inability to identify comparable sets of individuals within each group. Despite these motivations, the impact of ignoring such group influences when they exist (e.g. opting for single level or simple over complex multilevel) is unknown. In particular, it is informative to understand the tradeoffs in implementing a complex multilevel PS as opposed to a simple multilevel PS or single level PS. For instance, in assuming the true treatment assignment mechanism is a complex multilevel mechanism, theoretically employing a complex multilevel PS should reduce bias as compared to other PS's. However, as seen in the PS variable selection literature, the reduction of such bias in finite sample sizes may or may not be negligible (e.g. Brookhart et al., 2006). Further, such bias reduction is often coupled with an increase in variance. In particular, because the complex multilevel PS necessarily represents a more complex model in the sense that it utilizes more fixed and random effects, the variance of the corresponding treatment effect estimator can be inflated when compared to simpler models. In other words the bias reduction advantages of the multilevel PS are potentially met with variance inflation disadvantages representing the well known bias-variance tradeoff. As a result, in finite sample sizes, the various multilevel PS's may not represent a superior to that of a single level PS's.

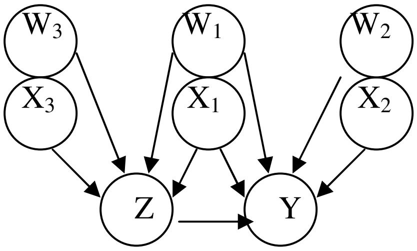
To assess the performance of the structures in multilevel settings in the context of different PS uses and types of variables, I performed four Monte Carlo simulation experiments. Whereas as prior work in multilevel PS's has focused more so on assessing multilevel PS through their performance through covariate balance, I shift focus to the treatment effect estimator (Kim & Seltzer, Hong & Raudenbush, 2006). Although the balancing property of the PS is a salient feature and diagnostic tool in empirical analyses, such focus in a simulation study where the true PS as well as other parameters are known is tangential. In particular, given the theoretical and previous simulation results as well as the design of this simulation study, models using the true PS score may achieve superior balance despite often producing inferior treatment effect estimators (Rubin & Thomas, 1996; Brookhart et al., 2006). In other words, the focus of this study is on how multilevel PS's influence treatment effect estimation rather than balance. Accordingly, to evaluate the various PS's in providing high quality treatment effect estimates, I followed single level PS literature and evaluated the PS's on the basis of the bias, variance and MSE of the corresponding treatment effect estimator.

To understand the role of such structures when estimating PS's, I examined the extent to which the each of the five multilevel PS structures posed influenced our estimates of the treatment effect. More specifically, I assessed the performance of each of the five structures in estimating the treatment effect when the true assignment mechanism was complex multilevel. Because the PS is a method to approximate ignorability of the treatment assignment rather than a treatment effect estimator, I combined the PS with matching, stratification or IPTW and an HLM outcome model to understand in practice how such choices affect an estimated treatment effect multilevel settings. As a result, I focused my inquiry on the performance of multilevel PS's implemented in HLM outcome models.

Further, to a large extent, the performance of the five multilevel PS's and three PS uses above in providing unbiased and efficient estimates depends on which types of variables are included in the PS. Accordingly, I extended the study to include three different types of variables based on prior literature: (1) those related to both the treatment and outcome; (2) those related to the outcome; and (3) those related to the treatment. In assessing the performance, I considered seven different variable combinations that one could include in the PS.

1. Only variables related to the both the treatment and outcome (1).
2. Only variables related to the outcome (2).
3. Only variables related to the treatment (3).
4. Variables related to the treatment and outcome plus those related to only the outcome (4).
5. Variables related to the treatment and outcome plus those related to only the treatment (5).
6. Variables related to only the treatment plus those related only to the outcome (6).
7. All available variables (7).

The first experiment examined the properties of the treatment effect estimator when using the different PS structures for a dichotomous treatment with other fixed parameters discussed below. Using each combination of the five different structures of PS models posed crossed with the three different PS uses and seven different PS variable specifications, I constructed PS models and then subsequently used them in HLM outcome models. The second experiment was similar, however, I assessed such performances among continuous treatments rather than dichotomous treatments. In addition, I limited the PS uses to stratification on the PS as matching or IPTW in the case of continuous treatments are not well studied. Figure (2.40) displays the simulated relationships.



Figure(2.40): Relation of variables: $X_.$ indicates individual level characteristics and $W_.$ indicates group level characteristics

In the third and fourth simulation experiments, I examined the sensitivity of the results from experiment one and two by altering the magnitude of the various relationships that seemed most relevant. These analyses were carried out by holding all other parameters fixed at their default values while a single parameter was altered. In particular, I consider nine different data generation parameters that may influence the performance of the PS's and their corresponding treatment effect estimators. The first parameter varied was the probability of receiving the treatment (*p)*. Specifically, the original experiment used a 0.5 probability of receiving the treatment whereas in the sensitivity analyses I used probabilities of 0.1 and 0.9. The second parameter I varied was the treatment effect (δ). In the original experiment I used a true treatment effect of 0.3 whereas second I used 0.1 and 0.5. Such values in educational data typically align with small, moderate and large effect sizes (Cohen & Cohen, 1988). Third, I varied the variance of the random effects from an original value of 0.2 to 0.1 and 0.3 (τ) as such values are typical in educational data (Coe & Hanita, 2009). Fourth I varied the correlation between the measured variables from a default value of 0.1, to 0 and 0.2 (ρ_x). Such values are fairly arbitrary although they generally represent weak to moderate relationships in social science data. Fifth I introduced a parameter that invokes a correlation between the random group effects on the treatment assignment and the random group effects on the outcome (ψ). Because prior literature in single level models has suggested that including variables minimally related to the treatment assignment but related to the outcome in the PS, this parameter allows us to understand how the relation of the outcome random effects and the treatment assignment random effects influences the treatment effect estimator. For example, as the bias reduction of a complex multilevel PS can be moderated an increase in variance, understanding the role of this parameter helps shed light on the bias variance tradeoff. In particular, I hypothesized that data with higher covariance between the outcome and treatment assignment random effects would increase the benefit of the using a complex multilevel PS in terms of both bias and variance. As a result, to understand the role of ψ I varied it from its default covariance of 0.01 to 0 and 0.05. As I am aware of no literature that considers such a parameter, the values and range have been subjectively chosen. Sixth, I varied the number of groups from the original value of 100 to 50 and 500 (n_j). Fifty groups was selected as a lower bound as it has been suggested that 50 groups is generally a lower threshold from which to effectively use multilevel models (Maas & Hox, 2005; Moinedden, Matheson & Glazier, 2007). Further, the upper bound of 500 groups was selected since few educational studies exceed this many schools. Seventh, in constructing the treatment assignment and the outcome I allowed the influence of each covariate to vary from a default of 0.5 to 0.2 and 0.8 (β). Again, though such values are fairly arbitrary, they represent typical effect sizes in educational literature. Next, I allowed the intra-class correlation to vary from a default of 0.2 to 0.1 and 0.3 (ρ_τ). Although such values align with those typically found in educational outcomes, such values for the treatment assignment are not well studied and highly dependent on the treatment (Coe & Makoto, 2009). As a result, I assumed these values as on the basis of educational outcomes and the subsequent application. Finally, in the case of stratification on the PS, I varied the number of subclasses from five to ten (*S*). Although, the number of strata is subjective and is often dependent on the common support in one's data, five has been a common though arbitrary number and is generally associated with removing about 90% of the bias (Rosenbaum & Rubin, 1983a).

All four simulation experiments employed the same data generating process. I generated the individual characteristics (X_1, X_2, X_3) and group characteristics (W_1, W_2, W_3) from two separate multivariate normal distribution with mean 0 and covariance matrices $\sum_{\mathrm{x}}$ and $\sum_{\mathrm{w}}$:

$$\sum\nolimits_{\mathrm{x}} = \begin{pmatrix} 1 & \rho_{\mathrm{x}} & \rho_{\mathrm{x}} \\ \rho_{\mathrm{x}} & 1 & \rho_{\mathrm{x}} \\ \rho_{\mathrm{x}} & \rho_{\mathrm{x}} & 1 \end{pmatrix} \qquad \sum\nolimits_{\mathrm{w}} = \begin{pmatrix} 1 & \rho_{\mathrm{w}} & \rho_{\mathrm{w}} \\ \rho_{\mathrm{w}} & 1 & \rho_{\mathrm{w}} \\ \rho_{\mathrm{w}} & \rho_{\mathrm{w}} & 1 \end{pmatrix} \tag{2.41}$$

In the dichotomous treatment case, I designed the treatment such that it represented the realization of a dichotomous variable given individual characteristics and group characteristics. In particular both the dichotomous and continuous treatment experiments allow the true treatment assignment mechanism to be a complex multilevel mechanism. For both dichotomous experiments, the true treatment assignment mechanism followed the hierarchical generalized linear model:

$$\begin{aligned} &Level1: logit(P(Z=1)) = \beta_0 + \sum_{i=1}^{3} \beta_i X_{ij} \\ &Level2: \beta_s = \gamma_0 + \sum_{j=1}^{3} \gamma_j W_j + u_s \end{aligned} \tag{2.42}$$

for *s* in 0, 1, 2, 3 and u_s~ multivariate normal with mean 0 and covariance matrix

$$\sum_{u_s} = \begin{pmatrix} \tau & 0 & 0 & 0 \\ 0 & \tau & 0 & 0 \\ 0 & 0 & \tau & 0 \\ 0 & 0 & 0 & \tau \end{pmatrix} \tag{2.43}$$

As depicted by Figure (2.40), the variables X_2 and W_2 as well as their interactions and random effects were constrained to be zero in (2.42). In the case of a continuous treatment experiments the treatment represented a normally distributed variable influenced both by individual characteristics and group characteristics. For both continuous experiments, the true treatment assignment mechanism followed the hierarchical linear model:

$$\begin{aligned} Level\,1: Z_{ij} &= \beta_0 + \sum_{i=1}^{3} \beta_i X_{ij} + \varepsilon_{ij} \\ Level\,2: \beta_s &= \gamma_0 + \sum_{j=1}^{3} \gamma_j W_j + u_{sj} \end{aligned} \tag{2.44}$$

where Z is the treatment, $\varepsilon \sim N(0,\sigma^2_{treatment})$ and $u_0 \sim N(0,\tau)$. Similar to the dichotomous experiments, the variables X_2 and W_2 as well as their interactions and random effects were constrained to be zero in (2.44).

Further the true outcome model for both experiments was a random intercept only hierarchical linear model (HLM)

$$\begin{aligned} Level\,1: Y_{ij} &= \beta_0 + \delta Z_{ij} + \sum_{i=1}^{3} \beta_i X_{ij} + \varepsilon_{ij} \\ Level\,2: \beta_0 &= \gamma_0 + \sum_{j=1}^{3} \gamma_j W_j + u_{0j} \end{aligned} \tag{2.45}$$

where Z is the treatment, $\varepsilon \sim N(0,\sigma^2_{outcome})$ and $u_0 \sim N(0,\tau)$ and the coefficients and random effects of X_3 and W_3 in (2.45) are constrained to zero according to Figure (2.40).

To estimate the treatment effect for each data set and approach in each experiment, I combined the three uses of the PS with a HLM (Hong & Raudenbush, 2006; Brookhart et al., 2006). Specifically, I constructed PS strata on the basis of quintiles and deciles of the logit of the PS and matches based on a full matching algorithm (Hansen & Klopfer, 2006). Further, in the dichotomous treatment case I constructed weights based on the inverse of the probability of receiving treatment (Robins, Hernan & Brumback, 2000). When using matching or stratification I utilized indicators and modeled the outcome as a HLM as follows:

$$\begin{aligned} &Level\,1: Y_{ij} = \beta_0 + \hat{\delta}Z + \sum_{i=1}^{q} \beta_q S_{qij} + \varepsilon_{ij} \\ &Level\,2: \beta_0 = \gamma_0 + u_{0j} \end{aligned} \tag{2.46}$$

where q is the number of strata or matches minus one. When using the IPTW PS, I utilize (2.46) without indicators and use weights. To compare the estimates based on the different model structures and specifications, I used the results of the Monte Carlo simulation experiment to estimate the bias and mean-squared error (MSE) of each approach. I estimated these quantities for a given approach using:

$$\widehat{Bias} = \frac{1}{M}\sum_{i=1}^{M} \hat{\delta}_i - \delta \tag{2.47}$$

and

$$\widehat{MSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} (\hat{\delta}_i - \delta)^2} \tag{2.48}$$

where M represents the number of simulated data sets.

Results

The results are presented in five general sections. The first section addresses the choice of PS model by contrasting the bias, variance and MSE of the five posed models. In particular, this section restricts contrasts to those estimates based on the same covariate specification and PS use. The next section addresses the role of covariate specification in the PS. That is, it only focuses on comparing which variables are included in the PS for a given PS model and PS use. The third section then focuses on contrasting the different PS uses for a given covariate specification and PS model. The fourth section then merges the first three by making comparisons across PS uses, covariate specifications and PS model choices. The final section then examines how the various sensitivity parameters above may influence the above results.

In this first section I compared the differences in PS model choice through the bias, variance and MSE of the corresponding treatment effect estimator. The first experiment I considered used stratification on the PS for a dichotomous treatment. Using the default parameters listed above, I contrasted the five model choices for a variety of variable specifications presented in Table (2.49). Evident from the rounded estimates in Table (2.49), the complex multilevel PS tended to illustrate the best performance in terms of MSE. Specifically, when the true confounders are included in the PS (e.g. $X_{1,} W_{1}$) the complex multilevel PS out performs the other PS model choices. However, when the true confounders were excluded from the PS the complex multilevel PS was regularly outperformed by other PS model choices. Similarly, in terms of bias, the complex multilevel PS tended to dominate the other model choices except when excluding the true confounders. Accordingly, one can see a certain bias-variance tradeoff in opting for more or less complex models. That is, while multilevel PS's tended to account for selection bias, they did so at a cost of inflated variance. Such a tradeoff makes it evident that relevant, high quality variables are most effective in constructing the PS.

Table(2.49): Rounded estimates of bias, variance and MSE for the dichotomous strata default experiment

Model Type	*10^2x*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
		X_1W_1	*X_2W_2*	*X_3W_3*	*$X_1X_2W_1W_2$*	*$X_1X_3W_1W_3$*	*$X_2X_3W_2W_3$*	*$X_1X_2X_3W_1W_2W_3$*
	Bias	4.12	18.97	19.42	1.73	1.70	17.91	2.75
Single	Variance	0.10	0.17	0.21	0.08	0.10	0.17	0.07
	MSE	0.27	3.77	3.98	0.11	0.13	3.38	0.15
	Bias	4.65	19.53	19.10	2.08	2.06	17.68	3.34
Single with fixed	Variance	0.10	0.18	0.21	0.07	0.10	0.17	0.07
	MSE	0.32	4.00	3.85	0.12	0.14	3.29	0.19
	Bias	3.82	19.29	19.12	1.37	1.36	17.62	2.45
Simple	Variance	0.10	0.18	0.21	0.07	0.10	0.17	0.07
	MSE	0.25	3.90	3.86	0.09	0.12	3.27	0.13
	Bias	3.53	19.23	19.25	1.79	1.77	17.72	2.07
Moderate	Variance	0.10	0.18	0.22	0.08	0.11	0.17	0.07
	MSE	0.22	3.87	3.92	0.11	0.14	3.31	0.12
	Bias	1.78	18.42	19.78	0.00	0.06	17.96	0.07
Complex	Variance	0.10	0.17	0.22	0.07	0.11	0.18	0.08
	MSE	0.13	3.57	4.14	0.07	0.11	3.40	0.08

*Estimates are multiplied by 100 and rounded off to two decimal points for ease of presentation. As such MSE may not be exactly equal to bias squared plus variance.

Shifting to the IPTW PS use in the dichotomous treatment case, I saw very similar trends. Though the relative ordering changed in certain instances, when including the true confounders, the complex multilevel PS regularly offered a lower MSE compared to the other models. Further, such models also tended to reduce bias the most. Table (2.50) presented the rounded estimates.

Table(2.50): Rounded estimates of bias, variance and MSE for the dichotomous IPTW default experiment

Model Type	*10^2x*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
		X_1W_1	*X_2W_2*	*X_3W_3*	*$X_1X_2W_1W_2$*	*$X_1X_3W_1W_3$*	*$X_2X_3W_2W_3$*	*$X_1X_2X_3W_1W_2W_3$*
	Bias	4.16	18.37	17.52	1.59	0.29	16.55	3.47
Single	Variance	0.10	0.20	0.23	0.08	0.11	0.20	0.09
	MSE	0.29	3.57	3.30	0.10	0.13	2.94	0.21
Single with fixed	Bias	3.12	18.37	16.78	-0.52	-0.47	15.81	2.42
	Variance	0.10	0.18	0.23	0.08	0.12	0.20	0.07

	MSE	0.20	3.56	3.04	0.08	0.12	2.70	0.13
Simple	Bias	2.65	18.30	16.97	-0.34	-0.40	15.98	2.27
	Variance	0.10	0.18	0.23	0.08	0.11	0.20	0.07
	MSE	0.17	3.53	3.10	0.08	0.11	2.75	0.12
Moderate	Bias	4.29	18.28	17.36	0.59	0.95	16.22	2.64
	Variance	0.11	0.18	0.22	0.08	0.12	0.19	0.07
	MSE	0.29	3.52	3.23	0.08	0.12	2.82	0.14
Complex	Bias	2.42	17.85	18.93	-0.01	0.27	17.20	0.24
	Variance	0.11	0.20	0.25	0.07	0.11	0.21	0.08
	MSE	0.16	3.38	3.83	0.07	0.11	3.16	0.07

*Estimates are multiplied by 100 and rounded off to two decimal points for ease of presentation. As such MSE may not be exactly equal to bias squared plus variance.

Similarly, matching on the PS for a dichotomous treatment and stratifying on the PS for a continuous treatment yielded similar results. However, to a small extent the advantage of using a complex multilevel PS over the others was dampened. That is, the reduction in MSE from a complex multilevel PS over the other models tended to be somewhat less on average. Table (2.51) presents the matching estimates whereas Table (2.52) presents the stratification on a continuous treatment estimates.

Table(2.51): Rounded estimates of bias, variance and MSE for the dichotomous matching default experiment

Model Type	10^2x	*(1)* X_1W_1	*(2)* X_2W_2	*(3)* X_3W_3	*(4)* $X_1X_2W_1W_2$	*(5)* $X_1X_3W_1W_3$	*(6)* $X_2X_3W_2W_3$	*(7)* $X_1X_2X_3W_1W_2W_3$
Single	Bias	-0.43	18.87	20.03	-2.52	-2.67	18.22	-2.17
	Variance	0.26	0.45	0.56	0.20	0.23	0.54	0.20
	MSE	0.25	4.01	4.56	0.26	0.30	3.85	0.25
Single with fixed	Bias	0.76	18.05	18.76	0.21	0.31	17.38	0.08
	Variance	0.24	0.41	0.60	0.16	0.20	0.49	0.17
	MSE	0.25	3.66	4.11	0.16	0.20	3.51	0.17
Simple	Bias	1.86	18.20	19.43	-0.03	-0.26	17.32	0.21
	Variance	0.20	0.45	0.55	0.18	0.23	0.51	0.16
	MSE	0.23	3.75	4.32	0.18	0.23	3.50	0.16
Moderate	Bias	1.69	18.08	19.34	-0.98	-0.52	17.42	-0.24
	Variance	0.19	0.45	0.58	0.17	0.22	0.52	0.18
	MSE	0.22	3.72	4.31	0.18	0.22	3.55	0.18
Complex	Bias	1.30	17.21	19.77	-0.09	0.10	17.51	0.25
	Variance	0.20	0.48	0.64	0.15	0.20	0.56	0.17
	MSE	0.21	3.44	4.54	0.15	0.20	3.62	0.17

*Estimates are multiplied by 100 and rounded off to two decimal points for ease of presentation. As such MSE may not be exactly equal to bias squared plus variance.

Table(2.52): Rounded estimates of bias, variance and MSE for the continuous stratification experiment default

Model Type	10^2x	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
		X_1W_1	X_2W_2	X_3W_3	$X_1X_2W_1W_2$	$X_1X_3W_1W_3$	$X_2X_3W_2W_3$	$X_1X_2X_3W_1W_2W$
	Bias	3.99	10.40	11.84	2.73	4.85	11.02	4.84
Single	Variance	0.01	0.02	0.04	0.02	0.03	0.03	0.03
	MSE	0.17	1.11	1.44	0.09	0.26	1.24	0.26
	Bias	3.98	10.81	11.76	3.30	3.54	11.01	3.56
Single With Fixed	Variance	0.01	0.03	0.04	0.01	0.02	0.03	0.02
	MSE	0.17	1.20	1.42	0.12	0.14	1.24	0.14
	Bias	3.84	10.77	11.80	3.14	3.05	11.04	3.07
Simple	Variance	0.01	0.03	0.04	0.01	0.01	0.03	0.01
	MSE	0.16	1.19	1.43	0.11	0.11	1.25	0.11
	Bias	3.50	10.75	12.26	2.51	3.16	11.44	3.17
Moderate	Variance	0.01	0.03	0.04	0.01	0.01	0.03	0.01
	MSE	0.13	1.18	1.54	0.08	0.11	1.34	0.11
	Bias	3.49	10.36	14.28	2.38	2.90	13.16	2.91
Complex	Variance	0.01	0.02	0.05	0.01	0.01	0.04	0.01
	MSE	0.13	1.10	2.09	0.07	0.10	1.77	0.10

*Estimates are multiplied by 100 and rounded off to two decimal points for ease of presentation. As such MSE may not be exactly equal to bias squared plus variance.

Despite the promising performance of the multilevel complex PS in each of the given simulations, this approach did not always dominate the others. For instance, the multilevel complex PS was particularly weak when including only variables related to the treatment assignment (e.g. combination (3) *X3,W3*) or those variables related only treatment and only to the outcome (e.g. combination (6) *X2,X3,W2,W3*). In other words if one can confidently identify confounding variables to enter in to the PS the complex multilevel PS may present a useful option. However, if one cannot confidently identify confounders, simpler PS model such as the single level with or without fixed group effects may offer an advantage.

Despite the MSE advantage of the complex multilevel PS over the other models with certain variable specifications, the absolute magnitude of such advantage in any case is relatively small. Specifically, the advantage of using the most appropriate model in any of the above circumstances is small relative to including the most appropriate variables.

When contrasting the seven different PS variable specifications, I noted the wide range in MSE among the different combinations. In particular, whereas the MSE's when examining PS model choice had relatively little fluctuation between models with the same variables (vertical columns in tables), variable selection (horizontal rows in tables) in the PS played a much larger role. For instance, for a given PS model such as complex multilevel, the estimates of MSE based on the seven PS covariate specifications ranged from 0.10 to 2.00 whereas within a fixed covariate specification the choice of PS model results in a much more constricted range of around 0.10 to 0.20. Such a pattern is evident in each of the simulations regardless of PS use or treatment type. Further, I saw PS's that excluded the true confounder (X_1, W_1) had considerably more bias and MSE than those that did not. In addition, I saw evidence that using variables that constitute the true PS's ((5) X_1, X_3, W_1, W_3) frequently resulted in higher MSE than other specifications. Specifically, PS models that included the true confounders and those only related to the outcome (e.g. (4) or (7) X_1, X_2, W_1, W_2) tended to produce the minimum MSE. Consistent with literature on non-nested outcomes, including all available variables may decrease the efficiency of the estimator and thus increases MSE (Brookhart et al., 2006).

Next I turn to comparing the three different PS uses within dichotomous treatments. Each of the three PS uses takes on a different but valid approach to adjusting for selection bias. In terms of MSE of the treatment effect estimator in multilevel models, stratification, IPTW and matching tended to have similar performances. There was some evidence that stratification and IPTW slightly outperformed matching, however such a difference was small. The differences in MSE between PS uses were generally comparable to the differences in MSE between PS models. In other words, relative to variable specification, both PS use and PS model choices changed the MSE little.

In looking at the results of each section holistically, there is evidence to suggest that when complex multilevel treatment assignment mechanisms exist, complex multilevel PS's may be more appropriate. However, such evidence is relatively weak in that the relative gain in terms of MSE in using such an approach is small. Further, such gains are increasingly small when compared to how much PS model variable specification contributes to MSE. For instance, for a dichotomous treatment in which we have stratified on the PS, adopting a complex multilevel PS with true confounders and those only related to the outcome, the average error of our estimate would be approximately 0.026. However, if we rather adopted a single level model with fixed group effects using the same variables our average error would go us to approximately 0.03. In other words using a complex multilevel PS over a fixed effects single level PS would reduce our average error by about 0.004 or 13%. In contrast, the average error rises much more when excluding relevant variables. For example, in adopting a complex multilevel PS that excludes true confounders but includes those variables related to the treatment only, the average error is about 0.19. However, using the complex multilevel PS with the true confounders and those related to the outcome only yields an average error of about 0.026. The difference in errors resulting from variable selection represents a much larger range than resulting from model choice. Similarly, although highly dependent on context, the PS use matters very little in comparison to variable selection as evident from the tables above.

Results of the next set of experiments, which examined the sensitivity of these results to different parameter specifications, indicated qualitatively similar findings. My assessment of the sensitivity of the above results to various parameters indicated that the results tended to be fairly insensitive to the parameters I varied. In particular, of the nine parameters, I noted that only four of them exerted even minimal influence within the range I examined. The first parameter that altered some of the above findings was the probability of receiving treatment. Specifically, only with the IPTW PS use did the results change. Though qualitatively similar to the original results, the absolute magnitude of the MSE when using IPTW on the PS increased when the probability deviated from 0.5. Despite the trimming the weights at the 5th and 95th percentiles, such a result is likely due to the overweighting of certain observations that empirically appear to have extremely high or low probabilities. Also, such sensitivity is may be due to the sensitivity of the logit link to misspecifications. A potentially more robust link when using IPTW may be the robit link (e.g. Gelman & Meng, 2005). Next, increasing the number of strata used when stratifying on the PS tended to universally decrease MSE. Because one generally increases the homogeneity within strata when using more subclasses, such comparisons tend to remove bias although at a decreasing return.

One parameter that did tend to generate noticeable influence over the MSE of the treatment estimator, was the magnitude of the variable coefficients (β). In particular, decreasing the size of β tended to reduce the separation in MSE between models. That is, when variables contribute little to inferring treatment assignment the difference among multilevel PS's shrinks considerably. Likewise, increasing β tended to expand the separation in MSE between models. In other words, when the imbalances among treatment groups are relatively large, utilizing a multilevel PS tends to reduce the MSE noticeably. In a similar vein, increasing the amount of variation the group is responsible for in deciding an individual's treatment assignment tends to have some influence. When predicting the treatment assignment, increasing the variation at the group level (τ) tends to separate PS model choices. More specifically, the benefit of using a multilevel PS increases as τ increases. Table (2.53) provides an overview of the sensitivity analysis results by presenting the MSE for the fourth variable PS specification which includes both true confounders and those variables only related to the outcome.

Table(2.53): Sensitivity of results to parameters (MSEx100)

		Dichotomous	Dichotomous	Dichotomous	Continuous
		Strata	IPTW	Match	Strata
		V4	V4	V4	V4
Original	Single	0.15	0.21	0.26	0.09
	Single with fixed	0.19	0.13	0.16	0.12
	Simple	0.13	0.12	0.18	0.11
	Moderate	0.12	0.14	0.18	0.08
	Complex	0.07	0.07	0.15	0.07
P=0.1	Single	0.07	0.54	0.21	NA
	Single with fixed	0.12	0.35	0.17	NA
	Simple	0.11	0.18	0.18	NA
	Moderate	0.08	0.24	0.15	NA
	Complex	0.07	0.17	0.15	NA
P=0.9	Single	0.07	0.4	0.24	NA
	Single with fixed	0.12	0.25	0.19	NA
	Simple	0.11	0.19	0.18	NA
	Moderate	0.08	0.28	0.16	NA
	Complex	0.07	0.2	0.15	NA
Strata=10	Single	0.02	NA	NA	0.02
	Single with fixed	0.03	NA	NA	0.03
	Simple	0.03	NA	NA	0.03
	Moderate	0.02	NA	NA	0.02
	Complex	0.02	NA	NA	0.02
Effect=0.1	Single	0.17	0.28	0.15	0.08
	Single with fixed	0.22	0.18	0.14	0.12
	Simple	0.16	0.18	0.12	0.11
	Moderate	0.13	0.2	0.15	0.08
	Complex	0.09	0.09	0.14	0.07
Effect=0.5	Single	0.16	0.25	0.15	0.07
	Single with fixed	0.2	0.15	0.18	0.12
	Simple	0.13	0.14	0.14	0.11

	Moderate	0.12	0.16	0.16	0.08
	Complex	0.06	0.06	0.15	0.07
Tau_u=low	Single	0.17	0.23	0.28	0.08
	Single with fixed	0.19	0.19	0.23	0.13
	Simple	0.14	0.17	0.19	0.12
	Moderate	0.13	0.18	0.2	0.09
	Complex	0.08	0.09	0.15	0.06
Tau_u=high	Single	0.15	0.3	0.24	0.13
	Single with fixed	0.22	0.17	0.14	0.11
	Simple	0.15	0.16	0.15	0.1
	Moderate	0.12	0.17	0.14	0.08
	Complex	0.06	0.06	0.14	0.06
Rho=low	Single	0.13	0.24	0.2	0.08
	Single with fixed	0.15	0.15	0.15	0.09
	Simple	0.12	0.14	0.19	0.08
	Moderate	0.1	0.15	0.15	0.07
	Complex	0.08	0.07	0.15	0.05
Rho=high	Single	0.16	0.21	0.19	0.13
	Single with fixed	0.21	0.14	0.19	0.15
	Simple	0.15	0.13	0.2	0.13
	Moderate	0.12	0.14	0.18	0.1
	Complex	0.09	0.07	0.16	0.11
Psi=low	Single	0.15	0.22	0.23	0.1
	Single with fixed	0.19	0.16	0.14	0.12
	Simple	0.12	0.15	0.15	0.11
	Moderate	0.12	0.16	0.15	0.08
	Complex	0.07	0.07	0.15	0.07
Psi=high	Single	0.16	0.22	0.19	0.09
	Single with fixed	0.2	0.13	0.14	0.12
	Simple	0.14	0.12	0.15	0.11
	Moderate	0.11	0.15	0.16	0.08
	Complex	0.06	0.06	0.19	0.07
N=50	Single	0.17	0.21	0.26	0.09
	Single with fixed	0.19	0.17	0.19	0.12
	Simple	0.15	0.16	0.16	0.12
	Moderate	0.15	0.12	0.18	0.09
	Complex	0.1	0.08	0.18	0.07
N=500	Single	0.14	0.15	0.19	0.09
	Single with fixed	0.15	0.15	0.18	0.1
	Simple	0.11	0.12	0.15	0.09
	Moderate	0.12	0.13	0.15	0.07
	Complex	0.06	0.07	0.13	0.06
Beta=0.2	Single	0.08	0.07	0.14	0.02
	Single with fixed	0.09	0.07	0.14	0.03
	Simple	0.08	0.06	0.14	0.03
	Moderate	0.08	0.06	0.13	0.02
	Complex	0.07	0.06	0.15	0.02
Beta=0.8	Single	0.38	0.43	0.21	0.11

	Single with fixed	0.35	0.38	0.18	0.19
	Simple	0.24	0.34	0.18	0.18
	Moderate	0.22	0.29	0.16	0.11
	Complex	0.07	0.06	0.14	0.1
Tau=0.1	Single	0.19	0.3	0.20	0.06
	Single with fixed	0.23	0.21	0.17	0.1
	Simple	0.18	0.2	0.24	0.09
	Moderate	0.15	0.26	0.22	0.07
	Complex	0.13	0.14	0.20	0.06
Tau=0.3	Single	0.26	0.26	0.35	0.17
	Single with fixed	0.2	0.16	0.11	0.13
	Simple	0.15	0.15	0.11	0.12
	Moderate	0.12	0.14	0.14	0.09
	Complex	0.06	0.05	0.12	0.06

Discussion

Though the results are specific to the parameters and data considered, they primarily suggested that researchers should consider multilevel PS's when it is likely the treatment assignment is influenced by group characteristics. Though there is a certain potential for loss in efficiency in accounting for such differences when they do not exist, this loss tended to be dominated by the potential bias added by ignoring the multilevel structure. As a result utilizing multilevel PS's tends to slightly improve the quality of the treatment estimator in terms of bias and MSE. However, the advantages of adopting a multilevel PS were moderated by several factors. First, the reduction in bias tended to noticeably exceed the increase invariance only when a complex multilevel PS was used. That is, in a number of circumstances the loss of efficiency tended to dominate the bias reduction in the simple and moderate multilevel PS's when compared to the single level models. Only when we fully adjusted for the multilevel nature of the treatment assignment did we see the reduction in bias dominate the increase in variance. More significantly, the advantages of using a multilevel PS are often moderated by the type of variables one has to include in the PS. In particular, if one does not have true confounders, those related to both the treatment and the outcome, one should consider adopting simpler PS models. Further, the results suggested that variable selection in the PS construction plays a much larger role than the type of model used. As is evident from the data above, the contribution of the PS model type is minimal compared to the contribution of variable selection. This study suggested that for the simulated data, MSE was generally minimized when we included those variables that were related to the outcome only in addition to the true confounders. The study also indicated that variables unrelated to the outcome but related to treatment assignment insert noise and make strata boundaries, matches or weights less clear. Such noise tended to increase variance without necessarily decreasing bias and thus tended to result in higher MSE. Such results are consistent with recent literature and suggest that using an approach that includes all the available variables or includes only those variables that predict the treatment assignment may decrease the quality of the estimator (Brookhart et al., 2006). Such results raise questions about the validity of common PS construction methods such as including all available variables or forward/backward stepwise variable selection. Such strategies exclusively focus on the treatment without consideration for the outcome and thus impair the ability of the PS to contrast appropriate groups based on the outcome under study. Examination of the results of sensitivity analyses provided evidence that the variable selection problem is complex in nested treatments and requires much further study. Evident from these analyses, variable selection played a much larger role that did any of the parameters considered. As a result, the centrality of advancing PS's may rest more on understanding the concurrent relationships that each variable possesses that in appropriate model selection. Further PS variable selection in real data analyses is much more complicated than the data generated in this study as multilevel settings pose a variety of complex situations and variable relationships not considered.

Application

Teacher quality has recently been a high priority in education literature, in part as a result of the No Child Left Behind Act (NCLB, 2001). This act requires states and schools to provide every student with highly qualified teachers and requires teachers to demonstrate subject matter competency through various means. In meeting these requirements states and schools have struggled to identify and understand the components of teacher quality and those attributes that are most representative of the components. Though definitions of teacher quality vary among states as well as within states, the core of teachers' competency is the specialized knowledge in the subject in which they teach. In the area of reading, indices of teachers' knowledge used to estimate teacher quality have primarily been indirect measures or "proxies" of knowledge (e.g., attainment of certification or an advanced degree in a reading-related area) rather than direct measures of knowledge about reading. However, research has not consistently demonstrated significant associations between such proxies of teacher knowledge and student achievement (Croninger, Rice, Rathbun, & Nishio, 2003). Research has considered qualifications such as attainment (e.g. Scholastic Aptitude Test) (Ballou, 1996; Ehrenberg & Brewer, 1995); years of teaching experience (Darling-Hammond, 2000; Hanushek, Kain, O'Brien, & Rivkin, 2005). Refined measures, such as number of courses or major, have sometimes been found to be related to students' academic gains (e.g., Croninger, et al., 2003), but not all studies have shown significant effects (Darling-Hammond, 2000; Goldhaber & Brewer, 2000). It is possible that more direct measures of teachers’ content knowledge in the area in which they are teaching could provide a better index of teacher quality.

In this application I focused on the extent to which teachers’ knowledge about early reading contributes to their students’ progress in reading. In particular, I assessed whether there is a direct measurable relationship between teachers’ knowledge about reading and students’ reading achievement, above and beyond that accounted for by other common teacher quality markers. That is, although attained credentials may exhibit some relationship with student literacy achievement, teacher’s literacy knowledge may additionally boost his/her effect on a child’s literacy. I included in my investigation a second question, and that is whether teacher literacy knowledge, as assessed by the current measure, is evenly distributed among schools. Specifically, as a byproduct of examining and controlling for teacher and school factors that influence teacher literacy knowledge, I assess whether teacher literacy knowledge is concentrated in certain schools.

In designing a study to determine the effects of teacher knowledge about reading on students’ reading acquisition, an important problem is identifying the types and extent of knowledge that teachers need to hold. This problem is particularly complex for teachers of beginning reading because, unlike mathematics and science, it is difficult to identify the content of instruction. Beginning reading instruction focuses on students’ acquisition of processes of word reading and comprehension. Teachers’ knowledge might be thought of as incorporating both an understanding of the process of reading and the methods by which children acquire skill in this process. A necessary aspect of the study, therefore, was developing a theoretical framework of teachers’ knowledge about early reading and designing and investigating the measure that reflected this theoretical framework (Kelcey, Carlisle, Rowan, Phelps & Johnson, 2008). In content domains other than reading, measures of teachers' knowledge have focused on pedagogical content knowledge, a construct that involves the intersection of the knowledge that teachers' need to impart and the methods used to convey this knowledge to students (Shulman, 1986). In contrast, theoretical frameworks and previous empirical studies of reading have focused on the extent to which teachers' hold linguistic knowledge about reading—that is, content knowledge. Because the framework and contents of measures of teachers' knowledge are likely to affect the outcome of a study of the measure as an index of teacher quality, I provide background in previous studies and discussion of the design of the study I carried out.

Rationale for Designing a Measure of PCK in Reading

In designing the current measure Kelcey et al. (2008) investigated the view that linguistic knowledge is the content knowledge that teachers of early reading must hold to be effective in teaching early reading. Such a view posits that teachers need to understand the linguistic structure of words to teach children to read; further, children need to understand how their oral language maps onto the written forms of words. In addition to placing emphasis on linguistic knowledge, the measure places emphasis on situated linguistic content knowledge. The premise is that teachers of early grades need to focus on not only a teacher's content knowledge but also their use of their knowledge in teaching reading. In contrast to proxies, a measure of content knowledge would allow researchers to explicitly assess a critical aspect of teacher quality. That is, in using proxies such as master's degrees, it may be unclear as to which skills gained through attaining a master's degree actually drive the effect of teacher quality or teacher knowledge on student literacy achievement (Kelcey et al., 2008).

Recent studies of teacher knowledge

Despite interest in assessing teachers' knowledge about reading, identifying such knowledge and examining its relationship to student achievement has remained inadequately specified in past research. Insufficient measurement of such knowledge has limited the causal conclusions one can draw. However, current interest in understanding and measuring teachers' knowledge and instructional capabilities has accelerated, and appropriate methods for studying the validity of such measures have improved. Included in such methods are efforts to examine the effects of professional development on teachers' knowledge and to examine the effects of teachers' knowledge on their practices and their students' reading.

The relation of teacher literacy knowledge and student achievement in literacy in the early elementary grades has been examined in relatively few studies. Those that do focus on such relationships have methodological features that make it difficult to infer the possible causality of teachers' knowledge about reading on their students' progress in reading. Two studies (Bos, Mather, Narr & Babur 1999; McCutchen, Abbott, Green, Beretvas, Cox, Potter, Quiroga & Gray, 2002) carried out assessments of teachers' knowledge about the linguistic foundations of reading, before and after teachers attended a program of professional development, but they then compared students' reading performance of teachers who did and did not attend the professional development program. Performance on the teacher knowledge measure did not figure into these analyses (Bos et al., 2001; McCutchen, Harry, et al., 2002). Other studies focused on students’ year-end performance rather than gains in reading over the year (Foorman & Moats, 2004). In addition, most of these studies did not directly assess the contribution of teachers’ knowledge about reading to their students’ gains in reading after taking into account other indices of teacher quality. Instead, the analytic strategy often focuses on unconditional comparisons. For instance, Foorman and Moats (2004; Moats & Foorman, 2003) compared teachers in terms of their attained teacher knowledge as part of a professional development program but without taking into account aspects of teachers’ professional background, such as degree attainment. In such circumstances, estimates can not be considered as causal effects of teacher knowledge, as the effect is entangled with other characteristics and may represent the summative effect of teacher quality.

To address this, one recent study by Cirino, Pollard-Durodola, Foorman, Carlson and Francis (2007), more rigorously assessed teacher knowledge by linking teachers with their students, as well as controlling for other teacher quality factors. Though this study found no effect of teacher knowledge on student achievement, its measure of teacher knowledge is a composite measure of teacher quality which includes a teacher knowledge component rather than directly considering teacher knowledge.

Second, those studies that do attempt to isolate the contribution of teacher knowledge to student achievement tend to ignore the influence of school and student characteristics. The information provided about the schools and students is sparse, and characteristics of the schools and students have often not been taken into account in the statistical analyses. Again, such estimates are difficult to consider as causal, because student, teacher, or school characteristics convolute the estimates. There is a strong likelihood that the larger context in which reading is taught affects student outcomes, as demonstrated recently by Kainz and Vernon-Feagans (2007). These researchers found changing patterns in the family and school factors that affected the development of reading skills from kindergarten through third grader.

With regards to student achievement, most studies find greater variance within classrooms than between classrooms or schools (Raudenbush & Bryk, 2002). As teacher knowledge is a teacher characteristic and is implemented to an entire class, the amount of variance it can account for is likely to be relatively limited. In addition, the complexity of the factors that affect students' reading outcomes and the difficulty of measuring teacher knowledge restrict my expectation that teachers' knowledge will account for their students' gains in reading across a year. Instead, a measure of teachers' knowledge might explain a rather small amount of the variance in students' reading gains across a year.

Evidence that modest expectations of the impact of the current measure of teachers' literacy knowledge on students' reading improvement is realistic comes from an earlier study (Carlisle, Correnti, Phelps, & Zeng, in press). The measure of teacher knowledge in this study focused on teachers' content knowledge about reading that was disseminated at the professional development seminars attended by teachers in Reading First schools in Michigan (Moats' *Language Essentials for Teachers of Reading and Spelling*; Moats, 2003). Using hierarchical linear modeling, the researchers controlled for socio-demographic characteristics at the student level and entered into the teacher level various characteristics of the teachers as professionals (e.g., certification status, educational attainment). The outcomes were the performances of first graders on two subtests of the Iowa Tests of Basic Skills, Word Analysis and Reading Comprehension, controlling for students' prior ability. The results showed no direct effects of teacher knowledge on students' gains in word analysis and reading comprehension.

The current study I report in this paper addresses the limitations of previous studies, the goal being a methodologically sound investigation of teachers' knowledge about reading. The study focuses on the knowledge about reading held by teachers in Reading First schools in Michigan. Reading First is Part B of the No Child Left Behind legislation, designed to improve the reading achievement of kindergarten through third-grade students in high poverty schools with chronic underachievement in reading. My primary research question is the extent to which first grade teachers' literacy knowledge effects student literacy achievement. The theoretical framework focuses necessarily on the question of the nature and extent of knowledge about reading that contributes to effective early reading instruction. Further, the design of the study intentionally takes into account the characteristics of schools and teachers that have not met these challenges successfully in the past. In this study I examined teachers' knowledge about reading by using their propensity to be a high-knowledge teacher and include various other indices of knowledge and experience (proxies such as educational attainments) so that interrelations of factors that influence teachers' knowledge acquisition are captured by the analysis.

Data

The data for this study were derived from the Reading First program in Michigan. Reading First is Part B of Title 1 of the No Child Left Behind Law of 2001. This legislation provides funding to support improvement of reading instruction in kindergarten through grade three in school districts with high levels of poverty and underachievement in reading. As part of this study, researchers collected demographic survey and student achievement data from student, teachers and schools in the entire 168 schools that participated in the Reading First program and evaluation for the 2006-2007 school year. To qualify for Reading First funding in Michigan, districts had to meet eligibility requirements of low reading achievement (i.e., 40% or more of 4th-grade students scoring below the proficiency cut point on the state assessment, Michigan Evaluation of Academic Performance, Reading; MEAP) for 2 of the preceding 3 years, and low income (e.g., 1,000 or more students from families below the poverty line). Of this state Reading First population, approximately 77% of grade one teachers volunteered to allow researchers to investigate the effects of teacher knowledge on student reading achievement. Collectively, the 373 volunteer teachers instructed over 5,720 students and were nested in 138 schools. Of those teachers who agreed to take part in research 297 grade one teachers had sufficient student data to be to be included in the analytic sample.

Although we were unable to conduct this study with the full population of Michigan Reading First teachers, we did have available data for both the population and research sample. This allowed us to compare the characteristics of the two groups to determine the extent to which the volunteer sample differed from the larger population of teachers. On nearly all measures, the two groups were minimally different. The only noteworthy difference between the two groups was on the measure of teachers' knowledge. On this measure, the volunteers scored significantly higher than the full population. Thus, with respect to teachers' knowledge, the research sample is not representative of all Reading First teachers in Michigan. Tables (2.54) to (2.56) compare the characteristics of the population who volunteered for research in the current study and the general state Reading First population.

Table(2.54): Characteristics of Reading First Research Students vs. Reading First State Student Population

	State Population N=9187	Analytic Sample N=5720
Disability	0.09	0.10
Limited English Proficiency	0.12	0.11
Special Education	0.04	0.04
Free or Reduced Lunch	0.73	0.71
Hispanic	0.11	0.12
White	0.36	0.40
Hawaiian	0.00	0.00
African-American	0.42	0.39
Asian	0.01	0.01
American Indian	0.01	0.01
DIBELS Fall NWF Score	22.56	23.73
Male	0.51	0.50
Average age (in months)	84.56	84.67
ITBS-Reading Comprehension	149.58	149.85
ITBS-Word Analysis	148.72	149.14

Table(2.55): Characteristics of Reading First Research Teachers vs. Reading First State Teacher Population

Characteristic	State population N=524	Analytic Sample N=297
White	0.76	0.83
African-American	0.16	0.10
Hispanic	0.04	0.07
Bachelors in Elementary Education	0.69	0.69
Bachelors in Early Childhood Education	0.07	0.09
Bachelors in Literacy Education	0.01	0.01
Bachelors in Special Education	0.15	0.15
Masters Degree	0.64	0.60
Masters in Elementary Education	0.33	0.58
Masters in Early Childhood Education	0.14	0.08
Masters in Literacy Education	0.16	0.14
Masters in Special Education	0.16	0.38
Post Masters Degree	0.06	0.05
Standard Certification	0.66	0.68

Provisional/Temporary Certification	0.20	0.21
Reading Certification	0.04	0.06
Special Education Certification	0.19	0.10
High years teaching	0.49	0.49
Number of Professional Trainings	3.55	3.74
Teacher New to Reading First in 2006-07 School Year	0.15	0.16
Average age of students in classroom	84.91	84.92
Average percent of male students in classroom	0.53	0.53
Average number of students in Special education	0.08	0.08
Average percent of disabled students in classroom	0.14	0.14
Average percent of limited English proficiency of students classroom	0.11	0.11
Average percent of students eligible for free or reduced lunch in classroom	0.72	0.72
Average percent of Hispanic students in classroom	0.11	0.11
Average percent of White students in classroom	0.41	0.41
Average percent of Hawaiian students in classroom	0.00	0.00
Average percent of African-American students in classroom	0.38	0.38
Average percent of Asian students in classroom	0.01	0.01
Average percent of American-Indian students in classroom	0.01	0.01
Average fall NWF/ORF of classroom	23.05	23.05

Table(2.56): Characteristics of Reading First Research Schools vs. Reading First State School Population

Characteristic	State population N=165	Analytical Sample N=138
Male	0.52	0.52
American Indian	0.01	0.01
Asian	0.01	0.01
African-American	0.50	0.46
Hispanic	0.12	0.13
White	0.35	0.38
Percent eligible for free or reduced lunch	0.74	0.77
Proportion of male teachers	0.07	0.07
Proportion of white teachers	0.79	0.78
Proportion of African American teachers	0.15	0.15
Proportion of Hispanic teachers	0.05	0.05
Proportion of Asian teachers	0.01	0.01
Proportion of teachers with bachelors degree in Elementary Education	0.69	0.71

Proportion of teachers with bachelors degree in Early Childhood Education	0.08	0.09
Proportion of teachers with bachelors degree in Literacy Education	0.01	0.01
Proportion of teachers with bachelors degree in Special Education	0.16	0.15
Proportion of teachers with any masters degree	0.61	0.61
Proportion of teachers with masters degree in Elementary Education	0.38	0.39
Proportion of teachers with masters degree in Early Childhood Education	0.08	0.08
Proportion of teachers with masters degree in Literacy Education	0.11	0.11
Proportion of teachers with masters degree in Special Education	0.30	0.3
Proportion of teachers with post masters degree	0.04	0.05
Proportion of teachers with standard certification	0.64	0.63
Proportion of teachers with provisional or temporary certification	0.23	0.24
Proportion of teachers with reading certification	0.05	0.04
Proportion of teachers with special education certification	0.14	0.12
Average number of professional trainings	3.45	3.49
Proportion of teachers with high number of years experience	0.46	0.45
Proportion of teachers new to reading first in 06-07	0.21	0.21

Measures of Students' Reading Achievement

The outcome measures for this study were the Iowa Test of Basic Skills (ITBS) standardized subtests concerning word analysis and reading comprehension published by Riverside Publishing. Word Analysis involves identifying and matching sounds and spelling elements of words. Reading Comprehension involves selecting responses to questions that followed short passages. The measure of students' performance was the developmental standard score (SS). According to information reported in the ITBS test manual, the median SS is 150 for first graders (The University of Iowa, 2003).

As reported by Riverside, the reliability (computed with Kuder-Richardson Formula 20) for each subtest for grade one is as follows: Word Analysis: .85; Reading Comprehension: .91. Content validity was established through designing a measure that corresponded to widely accepted goals of reading instruction in schools across the nation; the skills and abilities have been judged to be appropriate through a process that includes curriculum review, preliminary item tryout, national item tryout, and fairness review. Information on predictive validity was not available for grades earlier than fourth;

however for fourth grade, performance on ITBS significantly predicted performance on 12th-grade Iowa Test of Education Development (.68) and 12th-grade grade point average (.53) (Hoover, Dunbar, Fisbee, et al., 2003).

The measure of prior achievement is drawn from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment, DIBELS is a set of fluency measures of early reading skills used to assess elementary students' progress in reading. For this study, one subtest, Nonsense Word Fluency (NWF), was used to establish status in reading in the fall of the year. NWF entails decoding two- or three-letter nonsense words on a printed page; credit is given for the number of letters correctly decoded in 1 minute. Though an additional pretest is desirable, namely the ITBS tests from the prior year, these scores are unavailable as it is not administered in kindergarten. Technical information in a document downloaded from the DIBELS website provides information on alternate form reliability of DIBELS measures (Assessment Committee, 2002). For NWF, the median was .83 for first graders. In terms of validity (Assessment Committee, 2002) concurrent validity, for NWF concurrent validity for first graders who had also taken the Woodcock Johnson Readiness had a median of .51.

Teacher Measures

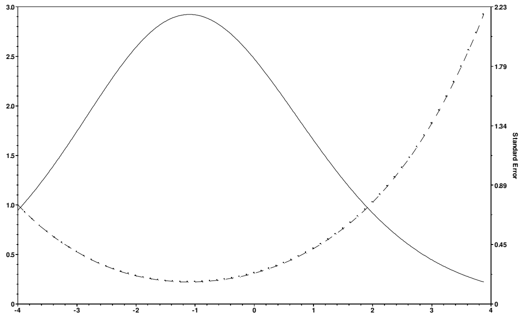
The primary source of data on teachers' backgrounds was the Teacher's Quest which was administered in the fall and winter of 2006 as well as the spring of 2007. This self-administered questionnaire focused on establishing measures of teacher experience, certification, education and professional training. Three sources of information came from this self-administered questionnaire: Language and Reading Concepts, Practices That I Use, and Teacher Information. The Teacher Information section gathered information about teachers' personal and professional characteristics such as race/ethnic background, undergraduate major, graduate major, attainment of a master's degree, type of certification they currently hold, and the type and amount of professional trainings they have attended. These factors were represented in a simple yes or no manner (represented through indicator variables) with the exception of professional trainings and workshops. For professional trainings, I constructed two measures of professional training. The first was a simple sum of the number of trainings completed by the teacher and the second was an indicator variable that separated those teachers that had completed more than the median number of trainings (median=3). Moreover, in summing the number of professional trainings, the measures excluded trainings listed as "other" as I could not validate the merit of such trainings. Acceptable professional trainings included the following options: Reading Recovery, Michigan Literacy Progress Profile (MLPP), LIFT, LETRS, Leading Professional Dialogues, Orton Gillingham, Spaulding, KLP (Kindergarten Literacy Profile), Six Traits Writing, Four Blocks, and DIBELS.

Language and Reading Concepts (LRC) is the measure of teacher knowledge and was developed as part of the Evaluation of Reading First in Michigan to be aligned to the professional development program in 2003-2005—*the LETRS program* (Moats, 2003; Kelcey et al., 2008). LETRS provides research-based lessons in the language basis for reading instruction and is made up of nine modules. In Michigan's professional development program includes instruction in each of these modules. Overall, the teachers received about 3 hours of instruction and practices for each module. LRC was developed to assess teachers' knowledge of principles and information in the LETRS program. In 2004-2005, LRC consisted of three measures, one administered in the fall, one in the winter, and one in the spring. In the subsequent year, fall 2005, the best 17 items were taken from these three forms to make up a single measure of reading knowledge. In the current administration, fall 2006, based on psychometric and theoretical properties of previous administrations, the focus of the test has shifted towards reading comprehension content and pedagogical knowledge (Phelps & Schilling, 2004). These items consist of the following: eight items (3-10) focus on phonemic awareness, phonics, and spelling; the remaining 9 items (11-19) focus on syntax/grammar, semantics/vocabulary, comprehension and comprehension instruction (Kelcey et al., 2008).

Psychometric analysis of the LRC teacher knowledge measure was conducted using data from the full population of first grade teachers participating in Michigan Reading First. Using Item Response Theory (IRT) (Hambleton, Swaminathan, & Rogers, 1991) and the software program BILOG I investigated scale properties and to scored participants (Mislevy & Bock, 1997) using a one parameter (Rasch) models. The IRT scale properties were 0.69 for IRT reliability and -1.13 for the test information maximum.

The test information curve is presented in Figure (2.57). The LRC assessment for first grade teachers has an information maximum below an average ability, providing the greatest power for reliably distinguishing among teachers at an ability of level of over 1 standard deviation below the mean. The information curve for the first grade teachers falls rapidly from the maximum dropping below an information of 2 at a teacher ability level of roughly a half standard deviation above the mean.

Figure(2.57): LRC (Teacher knowledge) Test Information Curve and Standard Error



In assessing the effect of the teacher knowledge captured by the LRC measure independent of other factors, I included a measure of the practices teachers use as to compare teachers with similar practices. I utilize the self reported practices measure from the fall, winter and spring teacher questionnaires. The measure of instructional practice consists of 42 student activities in phonemic awareness, phonics, fluency, vocabulary, comprehension, and writing that teachers are likely to use in early elementary literacy instruction. At each of the three administrations, teachers were directed to mark all instructional activities in which their students participated during the past full week of school. The set of activities marked by teachers provided a snap shot of the combination of practices teachers emphasize in their literacy instruction, emphasizing the richness of activities in which teachers engage students. I created one parameter IRT scale scores for each administration using Bilog and estimated scores based on the full state population of teachers participating in Michigan Reading First. The final total score is the average of the three administration scale scores and represents an overall measure of the emphasis that teachers place on the 42 student activities.

School and District Characteristics

The measures of school characteristics used in this study were drawn from the Michigan Department of Education website (www.michigan.gov/mde). This source provided data that was independent of the Reading First surveys. From this source, I constructed five measures. First, I identified the student population based on the percent of students that were male/female in a school. Second, I created an index based on the percentage of each race category (white, African-American, Hispanic, Asian, American Indian, Hawaiian, other). Third, I established a proxy measure for the socio-economic status of a school through the percentage of students that were eligible for free or reduced lunch. Finally, I estimated the collective qualities of Reading First teachers and students by aggregating all teacher and student covariates up to the school level. For instance, I recorded the proportion of teachers holding a masters degree and average DIBELS fall score.

Missing Data

Missing data retains the potential to bias parameter estimates and is a common complication in observational studies. Analyses with such partial data require making strong analytical assumptions that are unverifiable and often violated (e.g. missing completely at random (MCAR), Rubin, 1976). Analyses of this sort have the potential to misestimate effects and lead to erroneous inferences since most statistical inference tools are based on complete data sets. Rather than remove those students or teachers that have incomplete data, I employed the multiple imputation to impute missing values (e.g. Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). Multiple imputation assumes data are missing at random (MAR), a, perhaps, more plausible assumption than the missing completely at random (MCAR) assumption of listwise deletion (Rubin, 1976).

In this study, approximately ten percent of teachers and students had at least one missing data point resulting from teacher/student mobility and/or absenteeism on the day the data collection instruments were administered. Missing data analyses did not detect a significant difference in the observed covariates for students with missing data when compared to students without missing data. Additionally, although unverifiable directly, the data suggest that student and teacher missing data were unrelated to achievement and teacher knowledge and I accepted the MAR assumption. To appropriately address this missing data issue, I employed multiple imputation. In each imputation procedure, I based the student imputations on the full state population and considered all available variables measured at all levels to increase the robustness of inferences to violations of the MAR assumption (Peugh & Enders, 2004).

Analyses

My analyses are separated into two independent and parallel strands, each of which focuses on a single outcome. The first strand examines the causal effect of teacher literacy knowledge on individual student achievement in word analysis, while the second focuses on reading comprehension. In both of the models, my strategy consisted of two stages to approximate these causal estimands. In the first stage, I used the observational data to approximate an experiment using extensions of the canonical PS presented above. In the second stage, I combined PS stratification with hierarchical linear models to estimate the causal effects of teacher knowledge.

Applying the Rubin causal model (1974), I defined the effect of teacher knowledge on a student within his or her classroom as the difference between outcomes at various levels of treatment (teacher knowledge). More formally, using the continuous treatment measure of teacher knowledge, I defined the set of potential outcomes for student *i* in classroom j and in school k as

$$\mathbf{Y}=\boldsymbol{Y}_{ijk}^{(\boldsymbol{T})}\text{, for } t_i \text{ in } \boldsymbol{(T)} \text{ where } i=\{1, \ldots, n\} \tag{2.58}$$

where *T* is the set of potential treatment (continuous teacher knowledge) values and $\boldsymbol{Y}_{ijk}^{(\boldsymbol{T})}$ is a random variable that maps a particular potential treatment, t_i, to a potential outcome (Imai & Van Dyk, 2004). A first main assumption is the stable unit treatment value assumption (SUTVA) (Rubin, 1990). In this study, this assumption precludes the possibility of interference between students in different classrooms and thus accepts that the effect of a teacher's literacy knowledge on students in his/her classroom is independent of whether another teacher has high or low literacy knowledge. A second assumption needed is strong ignorability of treatment assignment (Rosenbaum & Rubin, 1983a). In the current context this assumption implies that the distribution of the teacher knowledge does not depend on potential outcomes given the observed covariates. Though SUTVA may be a tenable assumption as grade one presents self-contained classrooms and students experience minimal interaction with other teachers and their respective students, the ignorability of the treatment assignment need may not be tenable without proper adjustment as this is an observational study. In particular, teacher knowledge is a characteristic that is potentially intertwined with many pretreatment conditions. In theory, the likelihood that a teacher has a high degree of literacy knowledge is likely associated with his or her demographic characteristics; professional training and education, etc. Additionally, a school's and/or district's selection of teachers with a high degree of literacy knowledge is potentially associated with school/district characteristics. To make the assumption of strong ignorability reasonable, I utilized the PS.

Propensity Score Model

The multilevel dataset contains a set of variables that, theoretically, may be predictive of teacher knowledge and therefore predispose certain teachers to high teacher knowledge. Though literature on measuring teacher literacy knowledge effectively is scarce, there is strong evidence in the educational literature that levels of knowledge (usually student knowledge) tend to be clustered and be influenced by school characteristics and membership (e.g. Raudenbush & Bryk, 2002). Moreover, literature has demonstrated the clustering of high quality teachers in more affluent schools and districts (Weiler & Mitchell, 1992). Such relationships and clustering likely influence the dispersion of teacher knowledge. As a result, modeling the treatment assignment mechanism properly requires thoughtful consideration of these relationships. That is, I must effectively identify and estimate the varying selection mechanisms by which schools and districts assign teachers to a treatment level (i.e. simple, moderate or complex multilevel treatment assignment mechanism). For instance, one can envision that a teacher’s education may be fairly predictive of teacher knowledge in a school with a heterogeneous population whereas education may be trivial in its influence in a school with a homogeneous population. Such a setting would necessitate estimation that considers the differences in teachers as well as schools. Failure to account for such individual and group differences in the selection mechanisms, when they exist, will tend to bias the treatment estimator. Further to address the continuous nature of the treatment, I constructed the PS via a HLM as described above. Preliminary diagnostic analyses supported linearity as a feasible approximation of the true score.

A primary advantage of using the PS based approach has been its potential ability to mimic randomization. However, this advantage is often tempered by the difficulty of correctly specifying the PS as the bias and variance of the treatment effect estimator strongly depends on the subset of observed variables included in the construction of the PS. Theoretical literature has suggested that inclusion of variables unrelated to the treatment assignment but related to the outcome is vital, as their inclusion tends to decrease the variance of the estimator without increasing bias (see Variable Selection study within; Rubin & Thomas, 1996; Brookhart, Schneeweiss, Rothman, Glynn, Avorn & Sturmer, 2006). Moreover, it has also been suggested to exclude variables that are related to treatment assignment but unrelated to outcome as they increase variance without an accompanying decreasing in bias. Though such suggestions provide clear theoretical guidance, applying such principles in practice can be difficult. Further the difficult of such considerations increases when treatment assignment is multilevel. When treatment assignment is multilevel the number of explanatory covariates, cross level interactions and potential random slopes may increase rapidly. Datasets with even an average number of covariates and moderately large sample sizes may run into estimation problems as a result of limited degrees of freedom. Consequently, estimating PS's that are multilevel in nature requires an effective and efficient estimation strategy for variable selection. As evident from the literature, a variable selection strategy based on maximizing the prediction of the treatment only will neglect variables that are related to the outcome but weakly related to treatment assignment.

Further in the current context, literature on measuring and defining teacher knowledge is relatively weak and unsupported and the current measure represents on going research. Consequently, I relied on empirical evidence for model selection. Because of the number of teachers, my approach to constructing the level one (teacher) portion of the PS model was to include all measured teacher level variables. However due to the limited degrees of freedom for higher levels, I utilized a variable selection method similar to the Variable Selection study contained within. Whereas the Variable Selection paper within focused on the outcome-covariate and treatment covariate relationships separately, the current PS was constructed using the product of these relationships. In particular, using the pretest as a proxy for the outcome, I included those covariates in the PS whose product was the strongest in absolute magnitude. Similar to the Variable Selection paper within, I quantified such relationships through weighted partial correlations where the weights were estimated by the error variance and group level variances (see Variable Selection chapter).

For both word analysis and reading comprehension, I considered a three level propensity model where teachers are nested in schools which are nested in districts as each level illustrates significant variance (Table (2.59)).

Table(2.59): Unconditional Variance Components for Propensity Model

Component	Variance
Residual (e)	0.753
Schools (r)	0.096*
Districts (u)	0.063*

* *Ch*i-squared test p-value is < 0.01

My propensity model indicated that whether the teacher was African-American, whether the teacher had a bachelor's degree in early childhood education and whether the teacher had a reading certification varied randomly (e.g. complex multilevel mechanism) Using the following models I estimated each teacher's propensity to have teacher knowledge (complete specifications and variable list can be found in Kelcey et al., 2008):

Propensity Model: $\pi_{0j} = \beta_{00} + \delta Z_j + \sum_{q=2}^{10} \beta_{0q} S_{qj} + r_{0j}$

Level 1 (Teacher): $$TK_{jkl} = \pi_{0kl} + \sum_{p=1}^{n=P} \pi_p X_{pjkl} + \varepsilon_{jkl} \quad (2.60)$$

Level 2 (School): $$\pi_{pkl} = \beta_{p0l} + \sum_{q=1}^{Q} \beta_{pq} W_{qkl} + r_{pkl} \quad (2.61)$$

Level 3 (District): $$\beta_{000} = \gamma_{000} + \sum_{t=1}^{T} \gamma_t W_{tk} + u_{00k} \quad (2.62)$$

Applying the continuous treatment PS model, I subclassified teachers with similar values of the fitted propensity using quintiles and assessed balanced. When the PS is correctly estimated the observed covariates tend to be balanced among different levels of the treatment. Though there are more promising tools to assess comparability or balance, I assessed the comparability of subjects after estimating and controlling for the PS using the linear model (Hong & Raudenbush, 2006; Imai & Van Dyk, 2004). As I do not have a binary treatment but rather a continuous treatment, I compared models that omitted the estimated PS with models that conditioned on the estimated propensity function (Imai & Van Dyk, 2004). In the bi-variate model which omitted the estimatedPS, observed teacher knowledge is regressed on each covariate, individually, and the corresponding t-value for the respective covariate’s coefficient is presented. For each covariate, I placed the PS adjustment at the teacher level and the covariate at its proper level and use the propensity model previously specified:

Bi-variate Model:

Level 1 (Teacher): $$TK_{jkl} = \pi_0 + \pi_1 X_{pjkl} + \varepsilon_{jkl} \quad (2.63)$$

Level 2 (School): $$\pi_0 = \beta_{00} + r_{0kl} \quad (2.64)$$

Level 3 (District): $$\beta_{00} = \gamma_{000} + u_{00l} \qquad (2.65)$$

In the tri-variate model teacher knowledge is regressed on the same respective covariate as well as adjusted for the PS.

Tri-variate Model:

Level 1 (Teacher): $$TK_{jkl} = \pi_0 + \pi_1 X_{pjkl} + \pi_2 PS_{jkl} + \varepsilon_{jkl} \qquad (2.66)$$

Level 2 (School): $$\pi_0 = \beta_{00} + r_{0kl} \qquad (2.67)$$

Level 3 (District): $$\beta_{00} = \gamma_{000} + u_{00l} \qquad (2.68)$$

where PS represents the propensity score. The models are identical except that, in the latter, I controlled for the estimated propensity in each regression. The initial t-statistics, expectedly, are approximately normally distributed and show major covariate imbalances between levels of teacher knowledge (Appendix C). The t-statistics resulting from conditioning on the PS, are much closer to 0 since the model conditioned on the PS (Appendix C). As a result, balance on observed covariates for the teacher level as well as higher levels on teacher variables can be reasonably assumed.

Further to assess common support or adequate overlap between the estimated PS densities I extended the binary treatment idea of common support to the continuous treatment case. In the dichotomous treatment case, common support first requires a lack of perfect separability (linear, curvilinear, or otherwise) between the control and treatment groups implying that the probability of being a treatment or control can not be 0 or 1. Accordingly, dichotomous common support also requires that the propensity to receive treatment is not perfectly correlated with the actual treatment received, implying that there was some randomness in the process. In practice with binary treatments, the adequate common support condition is ensured by the simple presence of both treatment and control subjects in each stratum. The presence of both treatment and control groups in each stratum ensures the variance of the treatment condition within each stratum is not equal to zero, whereas the absence of both conditions implies the variance within a stratum is zero. This assessment ensures that the propensity for treatment and actual treatment are not perfectly correlated. Analogous to the dichotomous treatment situation, the continuous treatment PS common support condition requires ensuring that each PS based stratum contains adequate observed treatment variability and that the predicted level of treatment is not perfectly correlated with observed level of treatment. Appendix C displays this variation in observed teacher knowledge within each stratum. Moreover, the inter-strata overlap of actual levels of teacher knowledge, evident from the overlap between all interquartile ranges with all other strata, is complete and sufficient (i.e. all strata interquartile ranges have some considerable overlap with all other strata interquartile ranges). This complete overlap strongly suggests common support and sufficiently indicates a considerable lack of perfect correlation between the predicted and observed levels of teacher knowledge. Consistent with these graphs, bi-variate the correlation between the estimated PS's and observed treatment levels was less than perfect: ρ=0.88.

Finally, to understand the potential roles of the different multilevel treatment assignment mechanisms, I estimated three different PS's. The first represents a more historical approach and ignores clustering of teacher knowledge. With this approach, which I call a single level PS, I estimated the propensity to be a teacher with high literacy knowledge to solely be a function of teacher covariates. In essence, this approach implicitly considers the selection process as fixed across all schools/districts. The second

PS was the simpler multilevel PS while the third was the complex multilevel PS. Thus the estimated PS's represented increasingly complex views of the selection mechanism. Estimating these alternatives has several theoretical and practical utilities. First, I used PS adjustment to mimic randomization by comparing those subjects whose multivariate distribution of pretreatment covariates is similar. In this manner the PS assists in breaking any relationship between the treatment selection mechanism and the potential outcomes to approach unbiased estimates of the treatment effect. As a result, PS's that more accurately model the true selection mechanism may approximate the properties of randomization better than those which do not. In the current context, although true treatment selection mechanism is unknown, we can examine the comparative fit of the PS models to the data. As the PS models represent increasingly parsimonious views of the selection mechanism, I take advantage of the nested nature of PS models and compared the models' deviances. Under likelihood theory the distribution of the difference in deviances between nested models under the null hypothesis (of no difference) should have approximately a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters. Examining such differences may inform us about the general fit of each model relative to the others and identifies which model most closely resembles the true selection mechanism. As a result, such direct statistical comparisons help suggest which model most closely mimics randomization.

Further, contrasting such PS's and their respective estimates has several practical informative utilities. First, as estimating multilevel PS's can be a complex process, it is informative to understand whether this added complexity offers plausibly more accurate estimates. Second, in the case where there is little difference between the estimates, it is informative to understand the potential drawback of using the more complex options. In particular it is informative to examine the loss of efficiency resulting from a multilevel propensity framework. Third, it can be reassuring to get similar estimates from PS's constructed in different ways.

Outcome Models

Similar to the simulations above, I combined the PS stratification with regression adjustment using a HLM (e.g. Hirano & Imbens, 2002; Kleyman & Hansen, 2008). Specifically, I utilized the end of grade word analysis and reading comprehension outcomes from the ITBS subtests to examine the effects of teacher literacy knowledge. At level one, I considered seven student covariates that are historically related reading achievement. The first, male (π_1), is an indicator of whether the student is male. The second, age, (π_2) is a continuous measure of students age in months. The third, disabled (π_3), is an indicator that specifies whether a student has a known learning disability. Fourth, I considered whether a student has limited English proficiency (LEP) (π_4) and fifth I included an indicator for children who are eligible for free or reduced lunch (π_5). Sixth, I considered an indicator of whether a student is white (π_6). Lastly, I entered a measure of prior achievement (π_7). At the higher levels, I rely solely on the PS strata indicators and an intercept random effect at each level to adjust for selection bias.

Sensitivity Analysis

In addition to estimating the causal effects of teacher knowledge on student achievement, I assessed the robustness of my inferences to the inclusion of an unmeasured variable via a sensitivity analysis (e.g. Rosenbaum, 1995). Such analyses describe the magnitude of the relationships of an unobserved variable needed to alter the original inference. I attempted to control for overt known bias through the PS stratification and assumed that any unmeasured covariates, $\boldsymbol{U}$, are independent of treatment assignment given the measured covariates. I constructed a sensitivity index from the set of observed measures to determine if the model estimates are significantly influenced by potential hidden biases resulting from unobserved covariates. For the each statistically significant teacher knowledge effect, I examined whether the estimates would be significantly altered by additional adjustments for a hypothetical unmeasured confounder. My approach to determining the robustness of my estimates conceptualizes finite list of variables as being a representative sample of potential confounding variables (Hong & Raudenbush, 2006). Accordingly, I examined the impact of omitting one of the measured potential confounders on the estimates of teacher knowledge. I use these adjusted estimates as an index to gauge the robustness of the teacher knowledge effect.

Hidden bias may originate from any level in multilevel analyses, however I limit my sensitivity analyses to level two to correspond with the level the treatment was administered. I started by assuming that there exists an unmeasured teacher level covariate, U, comparable to the measured covariates. The impact of the omission of U on the estimate of teacher knowledge is dependent on the differences in teacher knowledge for the levels of U, represented by Γ, and its relationship with the outcome, represented by Δ. This impact is then potentially modified by the relationship between U and the measured covariates. Specifically, the impact of U on the treatment estimate may be reduced or absorbed if U is correlated with measured covariates (Frank, Duong, Maroulis & Kelcey, 2008). However, in my analyses, I considered U to be uncorrelated with all measured covariates thereby assigning *U* the maximal impact. First I estimate the Γ relationship, that of the treatment's with each of the covariates, ***X***, such that

$$E(TK) \; = \; \gamma_{000} + [\sum_i \gamma_{00i}(W_i)] \; + \Gamma(X_k) + [\sum_j \gamma_{j00}(A)_i] \; + v_{00} + r_0 + e \qquad (2.69)$$

where $e \sim N(0, \sigma^2)$, $r \sim N(0, \tau_\pi)$ and $v \sim N(0, \tau_\beta)$; A_j, X_k and W_i are level one, two and three variables, respectively, in the propensity model. Then, I examined each covariate's, X_i, hierarchical relationship with the outcome, Δ, controlling for those covariates at level one considered in the original achievement model

$$E(Y) \; = \; \gamma_{000} + \Delta(X_k) + [\sum_j \gamma_{j00}(A)_i] \; + v_{00} + r_0 + e \qquad (2.70)$$

where $e \sim N(0, \sigma^2)$, $r \sim N(0, \tau_\pi)$ and $v \sim N(0, \tau_\beta)$; A_i and X_i the level one and two variables respectively. Using these two relationships, I constructed multiple hypothetical unmeasured random variables, ***U***. Using ***U***, the unmeasured confounder, and the original estimate of the treatment effect, *δ*, I created a new estimate of the teacher knowledge effect, δ^*, that takes in account the unmeasured confounder

$$\delta^* = \delta + \mathit{\Gamma}\,(\mathit{\Delta}) \qquad (2.71)$$

I then sequentially assessed the sensitivity of the treatment to the addition of each U individually. That is, I assessed how different my inferences would be when assuming

$$\mathbf{Y} \perp \mathrm{TK} \mid \mathrm{X}_{\mathrm{measured}} \qquad \text{vs.} \qquad \mathbf{Y} \perp \mathrm{TK} \mid \mathrm{U}_{\mathrm{k}}, \mathrm{X}_{\mathrm{measured}} \qquad (2.72)$$

Results

I present the results in four sections. The first section attends to the characteristics and distribution of high knowledge teachers among teachers, schools and districts. The next section describes the results of the model based estimates of the teacher knowledge effect and is followed by a section that explicates the sensitivity of such estimates to unmeasured characteristics through a sensitivity analysis. The final section contrasts estimates based on PS construction that ignores clustering.

Several teacher, school and district level characteristics present strong relationships to a teacher's level of literacy knowledge. In particular, I saw that a teacher's knowledge was related to his or her race, negatively related to new to Reading First status, class's racial makeup and classroom practice. In particular, those teachers that emphasized a breadth of activities over a few activities tended to perform lower. In contrast, a teacher's knowledge was positively related to his or her reading certification status and educational specialization. In particular, I saw that specializing in literacy education, whether at the master's or bachelor's level is associated with higher knowledge scores. Similar to those teacher characteristics that influenced knowledge, a school's student and teacher racial compositions, the proportion of teachers specialized in literacy education along with the proportion of teachers maintaining a reading certification illustrated strong relationships with knowledge. In addition, the average number of professional trainings of teachers within a school attend demonstrated a strong relationship. Finally, the characteristics of a district also played a prominent role in predicting its teachers' levels of knowledge. Paralleling prior levels, the racial composition of the teachers and students within a district illustrated the most association with knowledge. However, at this level, the percent of students eligible for free or reduced lunch as well as the average prior achievement levels were related to knowledge.

Overall, these relationships strongly influenced the construction of comparable teacher sub-groups. The PS I developed indicated that teachers with a propensity to have higher levels of teacher knowledge tended to differ on most characteristics when compared with those teachers who maintained a propensity to have lower teacher knowledge levels. Although multicollinearity may convolute interpretation in propensity models, I noted that high teacher knowledge teachers tended to be white, specialize in literacy education and be experienced in teaching. Schools that maintained teachers with higher levels of teacher knowledge tended to have a high percent of teachers that were white, low percent of students that were African-American and had higher levels of approved professional training. Finally, districts that retained high knowledge teachers tended to have a lower percent of students that were African-American, higher percent of teachers that were white, and a higher percent of teachers that had obtained a masters degree.

I hypothesized that the teacher literacy knowledge examined here may not be evenly distributed among schools/districts but rather clustered within some schools/districts. As there is little research concerning the distribution of teacher literacy knowledge among schools/districts, I employed the multilevel PS model to empirically assess and subsequently address the clustering of knowledge in schools/districts. My analyses indicated that approximately 11% of the variation in teacher literacy knowledge is attributable to the school level and 7% is attributable to the district level (Table (2.59)). Though there is little prior research to compare this with, this decomposition is suggests substantial clustering and is similar, in magnitude, to those components found in student achievement. Moreover, such estimates are based on a Michigan Reading First sample-a group of districts and schools that are, to some extent, more homogeneous than the entire population of U.S. schools. Such estimates supported the hypothesis that schools/districts unevenly attract and retain high literacy knowledge teachers.

Using PS adjustments to create comparable sub-groups of teachers, I examined the effect of teacher knowledge on reading achievement via the ITBS subtests considering word analysis and reading comprehension. Using fully unconditional models, I partitioned the variance in each student outcome into three components representing the variance among schools, the variance of teachers within school and the variance among students within classrooms. As is typical, the majority of variation in student achievement was attributed to the students within classrooms component. The variation in this component often represents the variation in factors such as natural aptitude, motivation, family support but also is inclusive of measurement error. I saw smaller, yet statistically significant, estimates of the variance at higher levels. I found that the estimates of teacher and school components were responsible for roughly 9% of the achievement variation each. Table (2.73) presents the HLM estimates of the variance components for both the fully unconditional models as well as the final models.

Table(2.73): Variance Components for Achievement Models

	Word Analysis			Reading Comprehension		
Component	Final Model	Unconditional Model	Intercept Reliability (λ)	Final Model	Unconditional Model	Intercept Reliability (λ)
Teachers (r)	0.08	0.10	0.67	0.06	0.08	0.61
Schools (u)	0.03	0.08	0.51	0.018	0.08	0.51
Residual (e)	0.57	0.81		0.60	0.87	

The achievement model estimates of the effect of teacher literacy knowledge on student literacy achievement are presented in Table (2.74). The reading comprehension subtest produced the only statistically significant finding. In particular, my achievement model estimates the effect of teacher knowledge on this subtest to be 0.096 ($p=0.008$) That is, holding all other factors constant, for a one standard deviation increase in teacher knowledge I observe approximately a 0.096 standard deviation gain in student reading comprehension achievement over a one year period. Subsequent analyses concerning the word analysis subtest indicated that teacher knowledge has a small positive effect on student literacy achievement but is indistinguishable from zero.

Table(2.74): HLM Achievement Results

	Word Analysis					Reading Comprehension				
Effect	Estimate	SE	T	DF[1]	P-value	Estimate	SE	t	DF[1]	P-value
Intercept (π_0)	0.03	0.03	1.02	137	0.31	0.02	0.02	1.13	137	0.26
Teacher Knowledge (β_{01})	0.04	0.03	1.19	291	0.24	0.10	0.04	2.69	291	0.01
Strata 1 (β_{02})	0.07	0.08	0.89	291	0.37	0.15	0.08	1.90	291	0.06
Strata 2 (β_{03})	0.03	0.06	0.44	291	0.66	0.04	0.06	0.69	291	0.49
Strata 4 (β_{04})	-0.04	0.05	-0.80	291	0.42	-0.05	0.06	-0.80	291	0.43
Strata 5 (β_{05})	-0.02	0.07	-0.25	291	0.80	-0.04	0.06	-0.62	291	0.53
Age in months ($\pi1$)	0.00	0.00	-1.65	3763	0.10	0.00	0.00	-1.37	725	0.17
Fall non-sense word fluency ($\pi2$)	0.03	0.00	37.14	5707	0.00	0.03	0.00	33.25	5707	0.00
Male Student ($\pi3$)	-0.07	0.02	-3.52	2956	0.00	-0.14	0.02	-6.27	5707	0.00
Student Eligible for Free or Reduced Lunch (π_4)	-0.15	0.03	-5.14	5707	0.00	-0.20	0.03	-7.09	5707	0.00
Disabled Student ($\pi5$)	-0.31	0.05	-6.56	3896	0.00	-0.22	0.04	-5.21	290	0.00
Limited English Proficiency Student ($\pi6$)	-0.08	0.05	-1.82	578	0.07	-0.12	0.04	-2.86	186	0.01
White Student ($\pi7$)	0.23	0.03	7.62	5707	0.00	0.22	0.03	6.52	5707	0.00

[1] Degrees of freedom (and all other statistics) are adjusted for multiple imputation process (Raudenbush & Bryk, 2002)

Collectively, the models naturally suggest that student level factors were the most significant predictors of reading achievement. Most covariates at the student level illustrated strong relationships with both outcomes, including prior achievement measures (i.e. fall non-sense word fluency). This measure predicted students who had scored higher on the measure tended to score higher on both reading achievement outcomes when compared with those students who scored lower on the measure. A Student’s race, eligibility for free or reduced lunch and disability also played prominent and typical roles in predicting both outcomes in all three grades.

For the significant reading comprehension treatment effect, δ, I regarded it as sensitive to the strong ignorability assumption if it's new estimate, δ^*, was not significantly different from zero. In particular, examining all possible $\boldsymbol{\delta}^*$, I considered δ to be sensitive if the omission of an unmeasured covariate, U, with magnitude equal to X (measured covariate) created enough bias to alter my inference ($p>0.05$) (Appendix B).

The reading comprehension sensitivity analyses indicated that my estimate of teacher knowledge is robust to a wide range of characteristics but illustrates sensitivity to an unmeasured confound at the magnitude of a class's average prior achievement measure. Such sensitivity is typical, as measures of prior achievement often demonstrate the strongest impact in achievement models (Bloom, 2005). However, there is strong evidence that the inclusion of additional covariates accounts for decreasing amounts of variance once the most predictive (e.g. pretests) are considered (Bloom, 2005).

Finally, I compared three different types of PS's: single level, simple and complex multilevel. Initially, as they are nested sub-models of each other, I conducted hypothesis tests comparing the fit of each of the three PS's. Hypotheses tests revealed that the fit of the multilevel complex PS was superior to both simpler representations ($p<0.01$). Furthermore, in comparing the multilevel simple with the single level PS, I found similar evidence to suggest that multilevel simple had a superior fit ($p<0.01$).

In contrasting the respective PS based estimates of teacher knowledge I saw noticeable differences. In particular, for reading comprehension and word analysis, in utilizing the multilevel complex PS rather than a single level PS, I saw an increase in the estimated effect by a factor of about 2.5. In a similar but less magnified manner, selecting a multilevel complex rather than a single level PS translated into an increase in the standard error by a factor of slightly less than 2 (Table (2.75)). I expected such increases in standard errors as multilevel models expand traditional single level models by including additional covariates and cross-level interactions. However, it is informative to understand the magnitude of the reduction in efficiency by using multilevel models to estimate PS's. The bias-variance tradeoff is an important exchange to consider when contrasting propensity models. Its importance lies in the construction of the counterfactual. In particular, a main purpose in adopting the multilevel framework for the PS is to reduce bias by estimating the counterfactual via comparable individuals *and* groups. However, if that reduction in bias is linked with a high loss of efficiency, a benefit of the multilevel PS is weakened. My results for this data suggest this is not the case. For this data, I saw that the potential reduction in bias dominates the potential loss of efficiency from adopting a multilevel complex PS. This result is well aligned with both theoretical and applied literature concerning PS (Rubin & Thomas, 1996).

Table(2.75): Summary of teacher knowledge effects based on different propensity models

	Effect Size (SE)	
Propensity Model	**Word Analysis**	**Reading Comprehension**
Multilevel Complex (Current)	0.041 (0.035)	0.096 (0.035)**
Multilevel Simple	0.042 (0.03)	0.047 (0.028)^
Single Level	0.016 (0.024)	0.038 (0.019)^

** p<0.01 *p<0.05 ^p<0.1

In comparing the reading comprehension estimates based on the multilevel simple and the single level PS's there is a small yet relative difference in the estimates. However, in examining the incremental change in estimates (single to multilevel simple to multilevel complex), the largest shift in estimation comes from adopting the multilevel complex. In this data, this may provide evidence that not only is a single level PS largely inadequate but also a multilevel simple PS is inadequate. In contrast, with word analysis see a noticeable change when comparing estimates based on the single level model to those of the multilevel simple or multilevel complex models but see virtually no change when comparing the multilevel simple with the multilevel complex. Finally, though I estimated the models with the same data, it is reassuring that all estimates are positive and (marginally) significant.

Conclusion

In summary, my analyses revealed mixed evidence of whether the measure of teacher knowledge affects student reading achievement. My results demonstrated small effect sizes, in which statistically one effect can plausibly be prescribed to chance. Although such evidence is contrasting, I speculated that the measure of teacher knowledge may be more appropriate for gains in reading achievement as it is aligned with such content. In particular, analyses of the preliminary teacher knowledge measure (i.e. 2004 and 2005) from prior studies (Carlisle, Correnti, Zeng, submitted) indicated that the measure was dominated by word analysis tasks such as phoneme comprehension. Methodologically, I saw considerable separation in estimates using different PS models. Assuming the selection processes do indeed follow a mechanism that varies between schools in the complex manner, it is evident that ignoring such mechanisms may bias the causal estimate considerably. Theoretically, given the treatment mechanism is truly multilevel in nature, by correctly specifying the structural and stochastic form of the PS one can achieve strong ignorability of the treatment assignment. As a result, an approximation of a randomized experiment is most faithfully achieved through multilevel PS.

Chapter III

Variable Selection in Propensity Score Models for Multilevel Settings

Introduction

In studies that examine the causal effect of some treatment on an outcome of interest, propensity score (PS) methods and the Rubin Causal Model have become standard techniques (Rosenbaum & Rubin, 1983; Holland, 1986). Unlike standard parametric methods that control for confounding in an outcome model, PS methods rely on a model of treatment assignment to adjust for confounding (Brookhart, Schneeweiss, Rothman, Glynn, Avorn & Sturmer, 2006). In many observational education studies, it is difficult for researchers to confidently identify all relevant sources of selection bias. In such cases researchers are faced with a wide array of pretreatment covariates that may plausibly influence the treatment assignment. As the inclusion or exclusion of such pretreatment covariates can strongly affect the subsequent bias and variance of the treatment effect estimator, a central issue facing researchers using PS methods is how to select variables to be included in the PS model. As a result, construction of the PS plays a critical role in estimating causal effects and often requires practical strategies merging both theoretical and empirical evidence.

As observational education data often offer a wide range of characteristics that may be related to the treatment assignment and/or the outcome, it is central to understand how the inclusion of such covariates in the PS model affects treatment effect estimation.

Though preliminary research has explored the role of covariates with various relationships in the PS model, such research is limited in that it does not explicitly attend to the properties of the treatment effect estimator. In particular, common approaches such as including every available variable in the PS tends to focus exclusively on a single property of the estimator such as bias. Though the bias of an estimator is central, sole focus on the bias may degrade other properties that are highly relevant when faced with finite sample sizes. For instance, with finite sample sizes, sampling variation plays a significant role in estimating treatment effects. Consider an estimator which uses a single person's treatment effect as the estimator of the average treatment effect. While such an estimator is unbiased it will tend to have high variance as the estimate will likely be dependent on which person is chosen as a result of sampling variability. Such an estimator is undesirable in that it does not balance the tradeoff between the bias and variance of an estimator. Further, even when taking into account such a tradeoff through consistent estimators, the balance between the bias and variance of the treatment effect estimator is still critical in research with finite samples. Additionally, in considering studies with group level treatments, effective sample sizes are often reduced as the sample size is based on the number of observations at the group level rather than that at the individual level. For example, in studying the effect of kindergarten retention policies on the average achievement of schools, though we may have a student sample size in the thousands, the sample size most relevant in determining the bias and variance of a group level treatment effect estimator is the number of schools. In addressing such balances, literature has indicated that including covariates related to the treatment but not to the outcome in PS models can increase the variance of the estimator without a corresponding reduction in bias of the estimator (Brookhart et al., 2006). Yet, it may be difficult to identify and exclude such variables as most variables will have non-zero empirical relationships with the outcome in finite samples. Furthermore, standard PS model building strategies such as using every available variable or forward/backward stepwise regression neglect such considerations.

Research Questions

In this study I developed a PS model building method to measure the quality of the treatment effect estimator in terms of mean-squared error (MSE). Since MSE can be written as

$$MSE = (\text{bias}(\hat{\theta}))^2 + \text{var}(\hat{\theta}) \tag{3.1}$$

I developed a method that explicitly attends to the tradeoff between the bias and variance of an estimator. Specifically, I developed a method that identifies thresholds for which the reduction in bias of the estimator from adding a covariate is surpassed by a corresponding increase in variance. Such an approach emphasizes the importance of both the bias and the variance of an estimator through the MSE. Using this framework, I asked seven questions in the context of hierarchical linear models (HLMs):

1. How shall we define a framework for variable selection in propensity scores models that explicitly balances the bias and variance of a treatment effect estimator?
2. How can we extend the framework in (1) to encompass hierarchical data?
3. How shall we define and identify meaningful thresholds in which adding a variable with certain relationships to the PS model generally decreases the quality of the HLM estimator?

4. Do different uses of the PS have different thresholds? In particular, how do the thresholds of stratifying, matching, IPTW and covariance adjustment on the propensity score compare within the contexts of HLMs?
5. What are the properties of treatment effect estimators based on propensity score models built in this manner?
6. How does this method compare with standard propensity score model building techniques such as stepwise AIC selection and including all available variables?
7. Using the framework developed, how can we construct an empirical PS model building method that balances the bias and variance of the treatment effect estimator?

To address the first question, I developed a method based on the concept of impacts (Frank, 2000). Specifically, I focused on the two relationships defined by an impact: a covariate's unique relationship with the treatment and that covariate's unique relationship with the outcome. Using these relationships, I developed a method to construct PS models which balances the bias and variance of the estimator. In particular, I used Frank's framework to identify thresholds in which adding certain covariates would only increase the MSE of the treatment effect estimator. Subsequently, I adapted this method to address nested outcomes and group level treatments and assess its properties in such contexts by comparing its properties to that of standard PS model building strategies. In addressing these questions, I considered and compared the thresholds of four common uses of the PS within a hierarchical linear outcome model framework: (1) Stratification on the propensity score, (2) Inverse-probability-of-treatment weighting, (3) Covariance adjustment on the propensity score and (4) Matching on the propensity score.

Theoretical Framework

Causal Inference Using Observational Data the Propensity Score

Observational data are collected from a wide array of existing situations in education. In such data a researcher makes no attempt to manipulate the situation generating the data. As a result, the fundamental problem with observational data is that students and schools choose their situations or treatments according to some criteria. Accordingly, in estimating causal effects, researchers must adjust for factors that led the individual or school to their choice that might also be correlated with the outcome of interest. To appropriately adjust for the factors influencing treatment choice or assignment, researchers have developed a variety of methods to support causal inferences using observational data.

Traditionally, researchers have used ordinary least squares regression (OLS) to study the impact of school resources on outcomes (e.g., Coleman et al 1966). In the cross-sectional case, the analyst relies on a specification such as,

$$Y = X\beta + Z\delta + \varepsilon \qquad (3.2)$$

where ***X*** represent the control variables the analyst is attempting adjust for, ***β*** are the corresponding coefficients of the control variables, *Z* is the treatment assignment with coefficient *δ*, and *ε* is the error which has a normal distribution with mean zero and variance σ^2. However, among other assumptions, such an approach assumes, that all confounding variables have been measured and that there is a specific parametric relationship between the treatment and the outcome. OLS tends to be sensitive to such omitted variable and parametric structure assumptions. Further, such assumptions rely heavily on extrapolation in that they frequently estimate the counterfactual by extending the regression line beyond the scope of the data. As a result, OLS estimates are likely inaccurate in a variety of settings.

In another approach, regression discontinuity relies on the existence of a treatment assignment rule or cutoff. In particular, in regression discontinuity designs participants are assigned to a treatment groups based solely on a particular pretreatment characteristic or set of characteristics. Those participants that are above the cutoff receive one treatment and those below the cutoff receive another. Such designs often supply inferences similar to randomized experiments in that the treatment assignment is known to be unrelated to all confounders except that which determined the treatment. Though such a design is formidable, in observational data, rarely does a single pre-treatment variable decide the treatment assignment and rarely is there such a sharp or even fuzzy cutoff such that all above a certain value received the treatment.

Another possible basis for causal inference in observational studies is an instrumental variable approach. Such an approach relies on the existence of an instrument or in other words a variable correlated with the treatment but uncorrelated with the outcome conditional on other covariates. This approach attempts to identify variables that would not be expected to independently alter outcomes but do so only through an endogenous measured variable. Though this approach offers unbiased estimation of causal effects when a number of assumptions are met, its utility on observational data is reliant on the existence of a high quality instrument. Consequently, its use is generally limited to situations in which there are identifiable and measured exogenous events, e.g. certain policy changes, that have no direct effect on the outcome but rather work exclusively through an endogenous variable.

A fourth possible basis for causal inference in observational data is the PS and is the focus of this investigation. In particular, the focus of this study is causal inference in observational studies when a high quality instrumental variable or pretreatment cutoff variable is unavailable and PS based approaches provide a reasonable approach. PS based methods conceptually attempt to identify and contrast similar units to estimate a causal effect. Though such an approach has considerable flexibility in that it does not require a type of quasi-experimental setting that prior approaches did, it requires other assumptions. For instance, it assumes that one can reasonably infer the treatment assignment from the measured covariates and that all covariates that influenced the treatment assignment were measured. Below I provide more detail.

Causal inference in observational studies with PS based approaches within the RCM conceptually attempts to mimic a randomized experiment in which the treatment assignment mechanism is known to be a function of measured pretreatment covariates ***X***. In assuming this approach with observational data, the primary task then is to construct the PS such that the potential outcomes are independent of the treatment assignment given the PS.

In the case of two treatments, the PS represents the conditional probability of assigning each experimental unit to the treatment of interest. That is, assuming two different treatment conditions where Z=1 represents the treatment of interest and Z=0 represents some control condition, the PS, $e(\boldsymbol{X})$, is

$$e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X}) \tag{3.3}$$

Accordingly, if adjustment on all or a subset of measured covariates is sufficient for unbiased estimation of the treatment effect, then so is adjustment on the PS (Rosenbaum & Rubin, 1983a). That is, treatment assignment is conditionally independent of potential outcomes given the PS if the treatment assignment is conditionally independent of the potential outcomes given the measured covariates. The PS acts as a unidimensional balancing score in which all the information relevant to balancing treatment assignment in $\boldsymbol{X}$ is extracted in *e(**X**).* As a result, conditioning on *e(**X**)* balances the distributions of $\boldsymbol{X}$ between the treatment and control groups and thus ensures the strong ignorability of treatment assignment assumption needed for causal inference in the RCM. Accordingly, units with similar PS values but different treatment assignments can serve as counterfactuals estimates for the missing potential outcome. In other words, the expected difference in the observed responses to different treatment conditions when the PS is held fixed is an unbiased estimate of the average treatment effect. In observational studies, though $\boldsymbol{X}$ have been observed, the PS is generally unknown and needs to be estimated from the data. Though we generally do not have the true PS in observational data, estimated PS's tend to operate like true PS's in that comparing units on an estimated but fixed score tends to balance covariate distributions between treatment groups (Rosenbaum & Rubin, 1983a). In particular, theory has suggested that use of the empirical PS is often more effective than use of the actual PS as it tends to remove empirical confounders or chance imbalances due to sampling variability (Robins, Mark & Newey, 1992; Rosenbaum, 1987).

Uses of Propensity Scores

The literature surrounding the use of PS's has proposed several alternative uses of PS's for causal inference. In particular, since the PS is a tool to identify comparable units by balancing the distribution of their pretreatment covariates, the PS needs to be used to contrast treatment levels. In this project, I consider four alternative uses of PS's: covariance adjustment, subclassification, matching and inverse probability of treatment weighting (IPTW).

Covariance Adjustment on the Propensity Score

A first use of the PS is to directly adjust for it using covariance adjustment in a parametric model. Though such use may be straight forward, it often requires a high degree of confidence in the specified PS model and often relies on extrapolation similar to linear regression. In particular, if there are nonlinear or non parallel response surfaces between the treatment groups, the average treatment effect may be misestimated. In general, the problem a researcher faces in this use is that the linear discriminant based on the observed covariates may not be a monotone function of the PS. For example, if the variance and covariance matrices in the treatment groups differ, then covariance adjustment using the PS can increase rather than decrease bias. As a result, literature has generally steered clear of this option and relied on alternative approaches. Though this use of the PS in isolation requires caution, it can often be combined with other uses of the score such as subclassification and then embedded in a parametric model to improve the robustness of treatment effect estimates.

Stratification on the Propensity Score

A second use of the PS is to separate the experimental units in to subclasses of similar PS values. Such division creates similar covariate distributions among the treatment groups within a subclass. As a result, within a subclass the observed responses of the control units provide a reasonable basis for inferring the counterfactual responses of the treatment units. For instance, stratifying on quintiles of the logit of the PS tends removes approximately 90% of the bias associated with the measured covariates (Rosenbaum & Rubin, 1983a; Cochran, 1968). However, the number of strata needed tends to be influenced by the homogeneity or balance of covariates within subgroups between the treatment and control groups.

Matching on the Propensity Score

A third use of the PS is to match experimental units on the basis of the PS. The intention with this use is to create comparable sets of treated and control subjects. Similar to stratification, matching units on the PS creates similar covariate distributions among the treatment groups. As a result, the matched control unit responses provide a reasonable basis for inference on the counterfactual responses for the matched treatment units. Though exact matching on the PS is optimal, approaches such as matching the nearest available neighbor are utilized to make inference tractable (e.g. Rosenbaum, 1989). In particular, this use is most appropriate when there is a large reservoir of potential control units available. Though greedy matching schemes such as nearest neighbor provide simple approaches, it may not be optimal in terms of minimizing differences within matches (Rosenbaum, 1993). An alternative algorithm which attempts to minimize such global differences within matches is full matching (Hansen, 2004; Rosenbaum, 1991). In particular, this algorithm uses network flow designs (Hansen & Klopfer, 2006) and results in the smallest average distance within matched sets and contains one or more subjects from each treatment group in each matched set. In this study, I focus on the use of full matching as implemented in the R package *optmatch* (Hansen & Klopfer, 2006) to study impact based PS's used for matching in HLMs.

Weighting on the Propensity Score

A fourth use of the PS is to weight by the inverse probability of receiving the treatment (Robins, Hernan & Brumback, 2000). In particular, weights are constructed for experimental units by first estimating their probability of receiving treatment and then weighting them by the inverse of the probability in a parametric or non-parametric procedure. Such an approach creates a pseudo-population for each treatment group through weighting by the inverse probability of receiving the treatment. Under strong ignorability, the weighted mean difference between the treatment groups is a consistent estimate of the average treatment effect. However, such an approach relies heavily on the estimated weights and is easily influenced by the estimation and parametric structure of both the propensity model and the outcome model when used. Consequently, to make this method more robust to extreme observations, researchers often down weight extreme probability observations so that they do not exert undue influence. For instance, extreme weights may by those below the 5th percentile or beyond the 95th percentile. Such extreme weights are then trimmed and given weights equal to the 5th or 95th percentile.

Combining PS and Parametric Outcome Models

In adopting one of the four above PS uses, researchers subsequently evaluate the average treatment effect by contrasting the appropriate outcomes. In doing this one may additionally utilize a parametric or non-parametric structure to model the conditional relationship between the treatment and the outcome (e.g. Rosenbaum & Rubin, 1983a; Cochran, 1973). When parametric structures are appropriate, research has demonstrated benefits from combining the PS with, for example, regression adjustment (e.g. Hirano & Imbens, 2002; Kleyman & Hansen, 2008). Moreover, Robins and Rotnizky (1995; Robins, Rotnizky & Zhao, 1995) demonstrated that as long as only one of the models, either that for the conditional mean of the potential outcomes given covariates, or that for the treatment variable given the covariates, is correctly specified, the resulting estimator will be consistent. Of particular interest to this study and educational research in general, is addressing the multilevel nature of many educational phenomena. In particular, because students within the same classroom share the same teacher and school, we would like our treatment effect estimator to take into account the lack of independence between students. Consequently, this study focuses on adjusting for imbalances through the four different PS uses above within the context of a parametric multilevel model (e.g. Correnti & Rowan, 2007). Such an approach combines the estimated PS with a standard parametric HLM to address the nonrandom treatment assignment and the multilevel nature of education using a linear approximation of achievement.

Observational Study Design & Model

In educational research and other fields, research data often have a hierarchical structure. That is, the individual subjects of study may be classified or arranged in groups which themselves have qualities that influence the study. In this case, the individuals can be seen as the first level of units in the study and the groups into which they are arranged are second level units. Indicated by the questions and focus of this study, I concentrate on building impact based PS models for use in multilevel outcome models. To address this nested structure I utilize a hierarchical linear model (HLM) (Raudenbush & Bryk, 2002). In HLMs each level of the nested structure is formally represented by its own sub-model.

For example, in a two level model where students are considered to be nested within schools, we can represent the level one student model as

$$Y_{ij} = \pi_{0j} + \sum_{p=1}^{P} \pi_p X_{pij} + \varepsilon_{ij} \tag{3.4}$$

where Y_{ij} represents an outcome such as math achievement for the i^{th} student in school j, π_0 is the average student score adjusted for the student variables, X, and the corresponding coefficients, π_p while ε_{ij} has a normal distribution with mean zero and variance σ^2. To link the students and schools, we can represent the school through a sub-model or level two school model as

$$\pi_{0j} = \beta_{00} + \sum_{q=1}^{Q} \beta_{0q} W_{qj} + r_{0j} \tag{3.5}$$

where β_{00} is the average adjusted achievement for school, β_{0q} is average effect of covariate, W_{qj}, on adjusted achievement and r_{0j} is the random effect of school j and has a normal distribution with mean zero and variance τ_π. These sub-models articulate relationships among covariates within a given level and, in turn, express how variables at one level influence relations occurring at other levels (Raudenbush & Bryk, 2002).

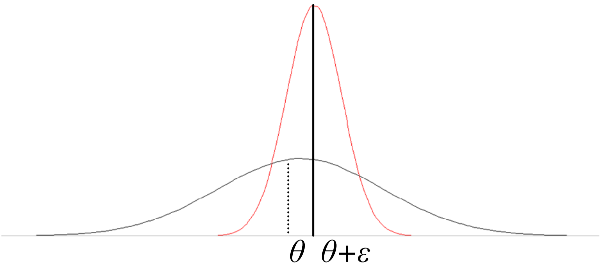
Though the general context of this study and scope of PS's focus on observational studies where treatment assignment was not controlled by the researcher, it is useful to outline the design of the randomized experiment one is trying to emulate using the PS and further adjustments. The method I discuss subsequently can be applied in a variety of settings and study designs, however, I focus this study on the HLM most closely resembling cluster randomized designs. In cluster randomized designs the experimental units are groups rather than individuals. Accordingly, intact groups are randomly assigned to treatments not individuals. For example, viewing schools as the grouping structure and students as the individual units, we may randomly assign fifty percent of schools and their respective students to a mathematics program that groups students by abilities and assign the remaining fifty percent of schools and their respective students to a non-ability grouping curriculum. Consequently the school level is the experimental unit as students in the same school provide the school's overall response to the treatment. Because the cluster level randomization ensures that treatments received by the students are independent of their potential outcomes, there is no need for random assignment of students to the clusters to obtain unbiased estimates of the treatment effect. Specifically, if the treatment, Z, is independent of the outcome, Y, and the school covariates, $\boldsymbol{W}$, then it is also independent of the student covariates, $\boldsymbol{X}$. Accordingly, cluster randomized designs are specifically considered when a treatment is inherently assigned at the school level. Such designs have become common in social science research as they facilitate inferences concerning group level treatments in a manner that acknowledges group constraints and dependencies of members within groups.

Though cluster randomized experiments ensure treatment assignment is ignorable, observational studies that reflect cluster randomized experiments may violate the ignorability assumption needed for the RCM. In particular, when treatments are assigned non-randomly at the cluster level, imbalances at the cluster level may insert bias into the treatment effect estimate. Moreover, though student characteristics co-vary with the outcome they are independent of the treatment assignment as treatments were assigned based on the characteristics of the entire group rather than of any individual. For example, a school may have selected to group their math students by ability because their respective abilities were highly variable. Here, students only influence the treatment mechanism in aggregate rather than individually. Consequently, the school level covariates that influence the treatment assignment, including the student level aggregates and their higher moments, are sufficient in removing all bias from the treatment effect estimator. Specifically, though ignorability is not met with unconditional comparisons of treatment groups, it is accomplished when we condition on (only) the school level factors. That is, the treatment assignment, *Z*, is independent of the student covariates, ***X***, and potential responses, *Y*, given the school characteristics, ***W***.

In addition to providing unbiased estimates, it is often of interest to use efficient estimators. Though omitting the student covariates, ***X***, in estimating the treatment effect will provide unbiased estimates, one can often more fully understand the within group processes and increase the precision of estimates by including the student covariates in the outcome model to explain some of the variation in the outcome. Excluding student covariates from the outcome model essentially ignores additional information and is impractical for applied researchers. To align with the goal of providing an effective and efficient estimator, I included group mean centered student level covariates in the first level of the outcome HLM. Their inclusion parallels more realistic analyses by allowing insight into group processes while increasing efficiency.

Variable Selection in Propensity Scores

A primary utility of the PS model approach is its potential ability to mimic randomization by making the treatment assignment conditionally independent of potential outcomes given the observed covariates. However, this utility is often tempered by the difficulty of correctly specifying the PS. In particular, the bias and variance of the treatment effect estimator strongly depend on the subset of observed variables included in the construction of the PS, especially in finite sample sizes. Consistent estimators ensure that the variance and bias of the treatment effect estimator goes to zero as the sample size approaches infinity. However, such consistent estimators in finite sample studies provide much less protection from the variance of an estimator as chance imbalances are much more likely. Relevant to education studies with hierarchical outcomes and group level treatments, such variability of the treatment effect estimator often plays a large role as the effective sample size depends on the number of groups rather than individuals. Consider Figure (2.18), in which, for a given sample size, two different consistent estimators are graphically compared. Though the first estimator, θ_1, is unbiased, its density is thoroughly dispersed throughout the parameter space indicating the estimator's variability. In contrast, the second estimator, θ_2, is slightly biased but its density is concentrated around its center. Consequently, we are forced to develop a criterion that accounts for both bias and variance in order to evaluate which estimator is more appropriate.



Figure(3.6): Density of two different estimators: Black is unbiased but fairly dispersed and red is slightly biased but concentrated around the estimand, θ

To attend to this tradeoff, Rubin and Thomas (1996) derived approximations for the reduction in the bias and variance of an estimated treatment effect using the PS. Such derivations support including all variables related to the outcome regardless of their relationship with the treatment assignment. Additionally, their derivations demonstrated that including variables that are strongly related to treatment assignment but unrelated to the outcome can increase the variance of the estimator without a corresponding decrease in bias. Accordingly, theoretical literature has suggested: (1) including variables unrelated to the treatment assignment but related to the outcome and (2) the exclusion of variables that are related to the treatment assignment but unrelated to the outcome as such an approach decreases bias without increasing variance (Rubin & Thomas, 1996; Brookhart et al., 2006). In other words, one should exclude those variables resembling the properties of an instrumental variable and if such a variable can be conceptualized as a high quality instrument consider an instrumental variable approach.

Though such suggestions provide clear theoretical guidance in PS model construction, applying such principles in practice can be complex. Although certain relationships may be theoretically hypothesized to be zero, empirically they may be nonzero. Furthermore, though bivariate relationships may be nonzero, multivariate relationships may approach zero. Consequently, Rubin (1997) suggests that variables related to the treatment assignment and theoretically unrelated to the outcome but which empirically demonstrate some nonzero relationship with the outcome, may be important to include. He argues that if such a variable had even a weak effect on the outcome, the bias inserted into the estimator by excluding the variable in the propensity model may dominate any potential loss of efficiency the estimator would experience by including the variable for a reasonable sized study (Rubin, 1997). In addition, Robins, Mark and Newey (1992) derived analytical results which demonstrated that the asymptotic variance of an estimator based on a potential treatment model is not increased and is often decreased as the number of parameters in the potential treatment model is increased. As a result, literature has suggested that the size of the PS model should increase proportional to the study sample size (Brookhart et al., 2006). In practice, such perspectives would suggest including many or all available variables, as variables rarely have zero relationships empirically.

In the context of education, classifying such associations between a covariate and an outcome is difficult as a result of the fundamentally multilevel nature of teaching and learning. For instance, in observational education studies that resemble cluster randomized trials, treatments are assigned to and enacted by schools as an entire unit rather than individual students. As a result, in estimating a covariate-outcome relationship, one must take into account the lack of independence with groups. To add to the variable selection problem, observational education data often offer a large and wide range of characteristics that illustrate non-zero relationships. It is central to understand how the inclusion of such covariates in the PS model affects the properties of the treatment effect estimator.

Though theoretical guidelines are emerging (Rubin, 1997; Brookhart et al., 2006), PS model construction in educational literature has generally relied on three approaches. The first such approach is stepwise selection. In this approach the treatment is initially modeled as a function of all available covariates. Subsequently, a researcher removes the covariate with the highest p-value greater than some critical p-value (e.g. 0.05). Next, we refit the model and remove the remaining least significant predictor provided its p-value is great than the critical value. This process is repeated until all covariates with p-values greater than the critical value are removed. Similarly, forward stepwise selection reverses the process by starting with no predictors and adding the most significant covariates sequentially. A third stepwise approach is to select a model based on an information criterion such as the Akaike (AIC) (e.g. Venables & Ripley, 2002). The stepwise AIC approach combines a forward and backward stepwise procedure to select a model that minimizes

$$AIC = -2ln(L) + 2p \tag{3.7}$$

where *ln(L)* is the log-likelihood and *p* is the number of parameters in the model. Another common approach, for observational education data sets with a sufficient number of sampled groups, is to estimate the PS using every available variable. Such an approach tends to privilege the bias of an estimator over its variance at all costs. In other words such an approach automatically selects the more dispersed estimator in Figure (2.18).

Though such strategies are often practical, constructing the PS model in the above manners impair the ability of the PS to contrast meaningfully comparable groups as they exclusively focus on the treatment without consideration for the outcome. Consequently, they neglect the duality of confounding by ignoring the effects of variables that are related to the outcome but weakly related to treatment assignment. Such neglect often results in increased treatment effect estimator variance without a corresponding decrease in bias. Similarly, although including a variable that has little to no relationship with the outcome but a strong relationship with the treatment does not bias the estimator, it can add substantial variance to the estimator. Though most studies use treatment effect estimators that are consistent, adding such variance in finite sample sized studies can detract significantly from the quality of the estimator. A method that uses the PS to remove bias in an effective and efficient manner rather than precisely predicting treatment assignment should result in a more robust and reliable estimate. To this end I develop a method to balance the bias and variance of the estimator.

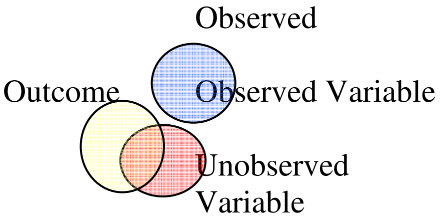
The Propensity Score as a Design and Analytic Tool

In a sense the PS can take on at least two different roles in facilitating causal inference. The first is as a design tool that solely provides balance of treatment assignments or randomization of treatments. In this capacity, one uses the PS solely to provide a reasonable basis for assuming randomization of treatments. That is, through PS uses such as matching, use of the PS as a design tool conceptually attempts to mimic randomization of the treatments. By identifying similar units based on their likelihoods to be assigned to different levels of the treatment, here the PS is used to devise a type of quasi-experiment. However, as a design tool, once the PS has provided a reasonable basis for such randomization through, for example, matches, the PS has no further involvement in estimating treatment effects. That is, using the PS as a design tool, one completely ignores the observed outcome processes and only utilizes the PS to devise a study that mimics randomized treatment assignment within groups with a common PS. Conceptually such use is similar to designing an experiment where one does not have access to the outcomes and randomizes treatments. Such an approach takes on more of a designed based approach in that it conceptualizes the sole role of the PS as a tool to make potential outcomes independent of the treatment assignment.

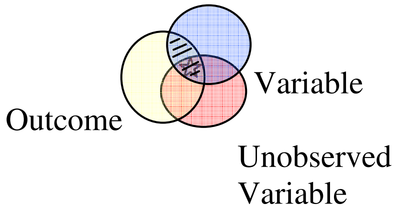
In a more integrated view, I consider the PS as both a design and analytic tool. In particular, though the PS is still used as a design tool to identify similar units, it takes an additional role as a tool for assessing treatment effects. In particular, as an analytic tool, the PS is used in the analyses of the outcome processes and their relations to the treatment. Such use helps facilitate identification of meaningfully comparable units rather than just simply comparable units. For instance, assume some treatment intends to improve mathematics achievement and is non-randomly assigned to students in a manner that depends on two characteristics of the students. First, suppose that students with low prior abilities are more likely to receive the treatment and second suppose students with blue eyes are more likely to receive the treatment. Further suppose, that we have no theory or evidence to suggest that the color of one's eyes will influence mathematics achievement, however we know that in this particular example the color of one's eyes did influence the treatment assignment. Using the PS exclusively as a design tool, we would holistically ignore the fact that, both theoretically and empirically, the color of one's eyes will have no effect on mathematics achievement. As a result, we will construct the PS model without any consideration of how eye color may influence mathematics achievement. Accordingly, we will then identify comparable students and, assuming we are matching on the PS, hope to match students who have similar prior abilities and similar eye color. Moreover, though two students have identical prior abilities, if they have different eye colors they will likely not be identified as comparable because they do not have the same eye color. In which case we may either ignore these students in estimating the effect or, more pragmatically, find other students who have slightly different prior abilities but the same eye colors and identify these students as better matches. As a result, the quality of our matches for the given outcome may be compromised because of our eye color restriction. In contrast, using the PS as both a design and analytic tool would indicate to the analyst that although eye color influenced treatment assignment it did not influence mathematics achievement. As a result, there is no need to include eye color in construction of the PS and we should identify comparable students on the basis of prior ability only.

Further including a variable such as eye color may not provide protection against unknown confounding variables. Specifically, matching on all characteristics relevant to the treatment assignment including, for example eye color, regardless of their relation to the outcome will not necessarily absorb or reduce the bias of an unknown confounding variable. That is, one might suggest that although eye color is not directly related to the mathematics achievement, it may be related to an unknown confounding variable and the inclusion of eye color in the PS may absorb some of the bias resulting from the omission of the unknown confounder's (Frank, Maurolis, Duong & Kelcey, 2009). This suggests eye color might act as a type of proxy for the unknown confounder. However, such absorption, or reduction in bias caused by an unknown confounder by the adjustment on a measured variable, is limited by the treatment-covariate and outcome-covariate relationships. Further, such absorption can be summarized by the product of the treatment-covariate and outcome-covariate relationships (Frank et al., 2009). Consequently if either of the treatment-covariate or outcome-covariate relationships is zero, no absorption or protection can occur. In other words, if eye color has no relationship with mathematics achievement (outcome-covariate relationship is zero) then it will offer no protection against an unknown variable which is confounded with the treatment effect (e.g. Figure (3.8)). In a similar manner, even if a variable does have some small relationship to the outcome and an unobserved variable, it can still only provide protection against the portion of the unobserved that it has in common which has already been accounted for. For example, in Figure (3.9) the observed variable can only account for the portion of the outcome-unobserved relation that is in common for all three variables (represented by a star).

Figure(3.8): Variable Relationships (1)



Figure(3.9): Variable Relationships (2)



Moreover, the inclusion of the observed variable depicted in Figure (3.8) (e.g. eye color) in the PS will decrease the efficiency of the corresponding treatment effect estimator without a corresponding reduction in bias. In this manner and in the identification and construction of meaningfully comparable units, we are not only respecting the duality of confounding but also using it to our advantage.

Using the PS as a design and analytic tool potentially identifies more meaningful comparisons for the process being studied and may do so in a manner that improves the MSE of our treatment effect estimator. Specifically, though the design approach may provide unbiased estimates of the treatment effect, it may do so in an inefficient manner as it requires the analyst to match on student characteristics that do not influence the outcome. Such an approach tends to increase the variance of the estimator by creating noise in the estimates. If we consider the case where an analyst must choose from hundreds of covariates, most of which frequently offer little to no information beyond core measure such as prior ability, we potentially add considerable amounts of noise and variability to our estimates.

If we accept the role of the PS to be both a design and analytic tool, we can potentially identify those covariates that truly influence both the treatment assignment and the outcome. However, such benefits are mediated by the potential to misestimate effects or their inferences as a result of using the observed outcomes in constructing meaningfully comparable groups. That is, the practical use of such an approach requires knowledge of the outcomes or each covariate’s relation to the outcome. Educational data that focus on the effects of interventions on achievement, however, have several features salient to this potential difficulty. Specifically, with regards to achievement, I address such difficulties in educational data by making use of a host of potentially available proxies. In other words, in order to construct meaningfully comparable groups while preserving the quality of inferences, we might consider using alternative measures of achievement to construct groups that are meaningfully comparable for the outcome of interest.

A first approach makes use of measures included in the current data that are generally known to be highly correlated with the outcome. In educational research where a post-test measures of achievement is the outcome, a measure of prior ability or pretest measure would often fit well in this approach. In this approach one would substitute the pretest for the outcome and construct the PS based on the proxy covariate-outcome relationships. A second similar approach is to use data for previous studies or earlier waves to estimate the outcome-covariate relationships. In this approach, for each potential PS variable, one would use data from similar study or prior wave to estimate the outcome-covariate relationships and inform PS construction. A third approach is to use cross-validation. In particular, when there a large amount of data exists, one can feasibly take a training or random sub-sample of the data and construct the PS based on the sub-sample's relationships. Subsequently, using the structural form determined in the training data, one would carry out the outcome analysis with the remainder of the data.

In a sense, utilizing proxies parallels a randomized experiment that utilizes covariates known to covary with the outcome to the increase power and precision of an estimator. If randomization was faithfully implemented, there should be little to no relation between the covariates and the treatment. However, because of the non-zero outcome-covariate relationships, adjustment on the covariates should decrease the variance of the treatment effect estimator. In this sense, rather than focus solely on an unbiased estimate, we utilize the outcome-covariate relationships as an analytic device to reduce the variance of the estimate and focus on an unbiased and efficient estimate.

Methods

Defining a Framework for Variable Selection

The method I propose makes use of the two central relationships each observed covariate has to summarize the bias that would be inserted by its omission in the PS model. The first relationship is that between the covariate and the treatment (Γ-relationship) and the second is that between the covariate and the outcome (Δ-relationship) (Gastwirth, Krieger & Rosenbaum, 1998). Drawing on Frank (2000), I quantify such relationships through sample (partial) correlations. Within the context of linear models, confounding can be demonstrated through measures of linear association such as correlation (e.g. Anderson, Auquier, Hauck, Cakes, Vandaele, Weisberg, Bryk & Kleinman, 1980). However, whereas Frank (2000) focuses on the product of these relationships, their use in the current context diverges. Although the product of these relationships summarizes the potential bias reduction of a variable on a treatment effect estimator, developing a PS model that exclusively focuses on this product may privilege the bias of an estimator over its variance. For example, such an approach would fail to discriminate between which relationship dominates the product. As a result, variables that are highly related to the treatment but minimally related to the outcome may be selected while those that are highly related to the outcome and minimally related to the treatment may be excluded. This would diverge with literature as it suggests including all variables related to outcome regardless of their relationship with the treatment and excluding those with minimal relation to the outcome. As the focus of this study is to develop a method that centers on bias and variance rather than bias alone, I retain the relationships defining an impact but do so using individual relationships rather than their product to understand the role each plays.

To balance the bias and variance of an estimator, I measured the estimator in terms of MSE. In particular, since MSE can be expressed as a function of the variance and bias of an estimator it provides a unique metric. Though there are numerous alternative criteria which may be used to judge the quality of an estimator, MSE is relevant in the current context for several reasons. First, since MSE is a function of the bias and variance of an estimator, it explicitly attends to the joint minimization of these quantities. Second, using MSE one can completely separate the contributions of bias and variance to the error of an estimator. This separation affords us explicit insight into how different types of variables contribute to the quality of an estimator. Third, in the current context, the multilevel treatment effect estimator is the maximum likelihood estimator which has an asymptotically normal distribution. Further, under a number of assumptions the minimum variance unbiased estimator for continuous normal distributed random variable is that which minimizes the squared error (e.g. Garthwaite, Joliffe & Jones, 2002). As such a particularly relevant criterion to judge the quality of an estimator is the mean squared error (MSE).

To see how MSE can be divided into separate components, we first write the definitions of bias, variance and MSE as

$$bias(\hat{\boldsymbol{\theta}}) = E[\hat{\boldsymbol{\theta}}\mathbf{]-}\boldsymbol{\theta} \tag{3.10}$$

$$var(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}})^2] \tag{3.11}$$

$$MSE(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] \tag{3.12}$$

where *E[·]* is the expectation, $\hat{\boldsymbol{\theta}}$ is the estimator, $\boldsymbol{\theta}$ is the estimand, and $\overline{\boldsymbol{\theta}}$ is the estimate of the first sample moment. We can expand the MSE of an estimator as

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}) &= E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] \\ &= E[\{(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}) + (\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})\}^2] \text{ where } \overline{\boldsymbol{\theta}} \mathbf{=} E[\hat{\boldsymbol{\theta}}] \\ &= E[(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}})^2] + E[(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] + 2(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})E[(\hat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}})] \\ &= var(\,\hat{\boldsymbol{\theta}}) \mathbf{+} [bias(\hat{\boldsymbol{\theta}})]^2 + 0 \end{aligned} \tag{3.13}$$

Using (3.13), we see that minimizing the MSE of an estimator explicitly translates into balancing the tradeoff between the bias and variance of an estimator.

Using these relationships in a manner similar to sensitivity analyses (e.g. Rosenbaum, 1986; Frank, 2000; Hong & Raudenbush, 2004), I examined the effect of the concurrent relationships on MSE thereby contrasting the potential reduction in bias with the potential increase in variance each covariate supplies. In particular, using the Γ- and Δ-relationships, I identified those variables whose inclusion in the PS will generally reduce the MSE of the corresponding treatment effect estimator.

Defining Thresholds

I defined these thresholds for a given treatment-covariate or Γ-relationship as the point in which the outcome-covariate or Δ-relationship causes the MSE of the treatment effect estimator to increase in comparison to its current MSE if such a covariate was added. Conceptually, such a threshold represents the point at which the reduction in bias (squared) is exceeded by an increase in variance in the treatment effect estimator. Though this method relies heavily on measures of linearity to summarize confounding, such measures are potentially representative of confounding in (hierarchical) linear models. As a result, bias tends to be reduced by inclusion of a covariate that has nonzero correlation with the outcome and nonzero correlation with the treatment. However, such reduction in bias is often met with an increase in variance thus emphasizing the need to balance the two properties. Alternative approaches that select covariates for the PS based solely on their relationship with the treatment, risk increasing the variance of the estimator without necessarily significantly decreasing bias. As a result, a PS model constructed based on such a threshold attempts to reduce the MSE of the treatment effect estimator. Though such thresholds would be most informative if they could be estimated from a given data set, such an approach is not generally possible as estimating the MSE requires prior knowledge of the true treatment effect. As a result, under some assumptions, I developed an approach for estimating thresholds through simulation to provide a model of how one might approach PS model construction.

Extending the Framework to Hierarchical Outcomes

Implementing an impact based PS model approach for use with a hierarchical linear outcome and group level treatment requires us to address two additional considerations first. The initial extension is identifying the two components of an impact as partial correlations rather than zero-order correlations. That is, I specify the Γ-relationship as the partial correlation between the logit of the treatment and the covariate of interest controlling for the other covariates and the Δ-relationship as the partial correlation between the outcome and covariate of interest controlling for the other covariates. This approach ensures the potential reduction in bias is not already accounted for. For instance, entering two variables that are co-linear may be redundant in terms of reducing the bias of the estimator. As a result, adding both would not necessarily decrease bias but likely increase the variance of the estimator. Though the need for this is evident in single level settings, in multilevel settings we must extend this to partial out relationships at both levels. Second, in the context of multilevel settings, complications are added as a result of the hierarchical nature of relationships. In particular, in estimating the relationship of a level two covariate with a level one outcome we need to consider random group effects and uneven sample sizes within groups. As a result, the estimated relationships between variables are likely to be more accurate in large groups than small groups thus creating unequal variance in the estimates. My approach capitalized on the approximation of maximum likelihood estimation in HLMs by weighted least squares

(WLS) when the variance components, σ^2 (error variance) and τ (group level variance), can be estimated and covariates are group centered (Seltzer, Kim & Frank, 2006). In particular, Seltzer, Kim and Frank (2006) show the weighted least squares regression based on cluster weights λ_j ,the aggregated level one covariates and level two covariates often provides an estimate extremely close to that of maximum likelihood in HLM (Seltzer et al., 2006). In such cases one should group center level one predictors and base the cluster weights, λ_j, on

$$\lambda_j = \frac{1}{(\sigma^2 / n_j) + \tau} \tag{3.14}$$

where σ^2 is the residual variance, τ is the between cluster variance and n_j is the number of level one units in group *j*. To address the hierarchical nature of educational data, I utilize a weighting scheme based on this ML-WLS estimation similarity to approximate magnitudes of the Γ- and Δ-relationships for each variable.

Using this framework, I assessed the effect of adding a covariate to the PS model on MSE using the respective unique hierarchical partial correlations. In other words, I consider the magnitudes of the partial weighted correlations such that these correlations represent the relationships exclusive of what other covariates (at all levels) in the model supply. Though the ultimate use of the PS will generally not be a simple covariance adjustment on the PS in the final HLM outcome model, the construction of strata and matched groups based on the PS will be considered through covariance adjustment via strata or matching indicator variables in the final HLM outcome model. Such an approach allows the PS to be a proxy for the respective indicator variables. Consequently a decrease in the bias of our HLM treatment effect estimator should manifest since confounding can be demonstrated by the correlation of a covariate with the outcome and the treatment in linear models (Frank, 2000; Anderson et al., 1980).

Identifying Thresholds

To develop such thresholds, I identified the MSE thresholds of covariates in the PS model by simulation and provide insight through derivation in the cases of the using covariance adjustment or IPTW on the PS in the outcome model.

To gain insight on the effect of including a covariate in a PS model, I considered estimating the treatment effect using the PS as a covariate in the HLM outcome model. Here, for simplicity, I assume we have a two level model where the treatment is at level two and is represented by an indicator variable, *Z,* the PS is entered as a level two covariate, *e(**X**),* and we do not consider individual variables. We write

$$\begin{aligned} &Level\,1: Y_{ij} = \beta_0 + \varepsilon_{ij} \\ &Level\,2: \beta_0 = \gamma_0 + \gamma_1 Z_j + e(\mathbf{X}) + u_{0j} \end{aligned} \tag{3.15}$$

where ε_{ij} has a normal distribution with mean zero and variance σ^2, and u_{0j} has a normal distribution with mean zero and variance τ. Moreover, I assume the WLS approach outlined above provides a good approximation of the ML parameter estimates resulting in the following model

$$\bar{Y}_j = \gamma_0 + \gamma_1 Z_j + e(\mathbf{X}) + \varepsilon_j \tag{3.16}$$

In estimating the parameters of (3.16) via WLS, γ_1 is equivalent to

$$\hat{\gamma}_1 = \left(\frac{s_{\bar{y}}}{s_z}\right)\left(\frac{r_{\bar{y}z} - r_{\bar{y}e(x)} r_{ze(x)}}{1 - r^2_{ze(x)}}\right) \tag{3.17}$$

where $r_{..}$ indicates the sample correlation, $s_.$ indicates the standard deviation and *e(**X**)* indicates the PS. In addition, the variance of $\hat{\gamma}_1$ is equivalent to

$$\mathrm{var}(\hat{\gamma}_1) = (\frac{s_{\mathrm{y}}^2}{s_{\mathrm{z}}^2}) * \frac{1-(\frac{r_{\mathrm{yz}}^2 + r_{\mathrm{ye(x)}}^2 - 2r_{\mathrm{yz}}r_{\mathrm{ye(x)}}r_{\mathrm{ze(x)}}}{1-r_{\mathrm{ze(x)}}^2})}{n\text{-}q\text{-}1} * \frac{1}{1-r_{\mathrm{ze(x)}}^2} \tag{3.18}$$

where *n-q-1* indicates the degrees of freedom. Without loss of generality I assume $\bar{Y}$, *Z* and the remaining variables have been standardized such that each has a variance of one and let *Y* replace $\bar{Y}$ for notational convenience. Given an estimand, *θ,* we can represent the bias of (3.17) by

$$bias(\hat{\gamma}_1) = E(\hat{\gamma}_1) - \theta \tag{3.19}$$

where we can estimate the expectation of γ_1 using the population correlations, $\rho_{..}$, with

$$E(\hat{\gamma}_1) = \frac{\rho_{\mathrm{yz}} - \rho_{\mathrm{ye(x)}}\rho_{\mathrm{ze(x)}}}{1-\rho_{\mathrm{ze(x)}}^2} \tag{3.20}$$

Accordingly the MSE of the treatment effect estimator can be estimated by

$$\begin{aligned} MSE(\hat{\gamma}_1) &= (E(\hat{\gamma}_1) - \theta)^2 + \mathrm{var}(\hat{\gamma}_1) \\ &= (\frac{\rho_{\mathrm{yz}} - \rho_{\mathrm{ye(x)}}\rho_{\mathrm{ze(x)}}}{1-\rho_{\mathrm{ze(x)}}^2} - \theta)^2 + \frac{1-(\frac{\rho_{yz}^2 + \rho_{ye(x)}^2 - 2\rho_{yz}\rho_{ye(x)}\rho_{ze(x)}}{1-\rho_{ze(x)}^2})}{n\text{-}q\text{-}1} * \frac{1}{1-\rho_{ze(x)}^2} \end{aligned} \tag{3.21}$$

Next, given two different PS's, e_ω***(X)*** and e_Ω***(X)***, let the first PS be constructed with *n-1* variables and the second be constructed with *n* variables of which the first *n-1* are the same. In general, we would like for the inclusion of an additional covariate in the PS to improve the *MSE(* $\hat{\gamma}_\omega$ *)*, indicating that

$$MSE(\hat{\gamma}_{e_\omega(X)}) > MSE(\hat{\gamma}_{e_\Omega(X)}) \tag{3.22}$$

Thus, we can estimate a threshold for the inclusion of the n^{th} variable by identifying the values for which the MSE of the estimator based on the first PS becomes equal to the MSE of the estimator based on the second PS. That is

$$MSE(\hat{\gamma}_{e_{\omega}(X)}) = MSE(\hat{\gamma}_{e_{\Omega}(X)}) \tag{3.23}$$

Consequently, using (3.21) we can obtain the threshold by setting the two equal

$$\begin{aligned}
&(\frac{\rho_{\mathrm{yz}} - \rho_{\mathrm{ye}_{\omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\omega}(\mathrm{x})}}{1-\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})}} - \theta)^{2} + (\frac{1-\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})} - \rho^{2}_{\mathrm{yz}} - \rho^{2}_{\mathrm{ye}_{\omega}(\mathrm{x})} + 2\rho_{\mathrm{yz}}\rho_{\mathrm{ye}_{\omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\omega}(\mathrm{x})}}{(n-q-1)(1-\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})})^{2}}) = \\
&(\frac{\rho_{\mathrm{yz}} - \rho_{\mathrm{ye}_{\Omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\Omega}(\mathrm{x})}}{1-\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})}} - \theta)^{2} + (\frac{1-\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})} - \rho^{2}_{\mathrm{yz}} - \rho^{2}_{\mathrm{ye}_{\Omega}(\mathrm{x})} + 2\rho_{\mathrm{yz}}\rho_{\mathrm{ye}_{\Omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\Omega}(\mathrm{x})}}{(n-q-1)(1-\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})})^{2}})
\end{aligned} \tag{3.24}$$

After reducing expression (3.24) (Appendix F), we can estimate the threshold in terms of ρ_{yz}, $\rho_{ye(\mathbf{X})}$ and $\rho_{ze(\mathbf{X})}$ using the quadratic equation

$$\begin{aligned}
&\rho^{2}_{\mathrm{yz}}\Big\{(\frac{(a-1)}{a(1-\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})})^{2}}) - (\frac{(a-1)}{a(1-\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})})^{2}})\Big\} + \\
&\rho_{\mathrm{yz}}\Big\{[\frac{-2(a-1)\rho_{\mathrm{ye}_{\omega}(\mathrm{x})}\rho_{\mathrm{ze}_{1}(\mathrm{x})} - 2a\theta + 2a\theta\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})}}{a(1-\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})})^{2}}] - [\frac{-2(a-1)\rho_{\mathrm{ye}_{\Omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\Omega}(\mathrm{x})} - 2a\theta + 2a\theta\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})}}{a(1-\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})})^{2}}]\Big\} \\
&+\Big\{(\frac{a\rho^{2}_{\mathrm{ye}_{\omega}(\mathrm{x})}\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})} + 2a\theta\rho_{\mathrm{ye}_{\omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\omega}(\mathrm{x})} - 2a\theta\rho_{\mathrm{ye}_{\omega}(\mathrm{x})}\rho^{3}_{\mathrm{ze}_{\omega}(\mathrm{x})} + a\theta^{2} - 2a\theta^{2}\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})} + a\theta^{2}\rho^{4}_{\mathrm{ze}_{\omega}(\mathrm{x})} + 1 - \rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})} - \rho^{2}_{\mathrm{ye}_{\omega}(\mathrm{x})}}{a(1-\rho^{2}_{\mathrm{ze}_{\omega}(\mathrm{x})})^{2}}) - \\
&(\frac{a\rho^{2}_{\mathrm{ye}_{\Omega}(\mathrm{x})}\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})} + 2a\theta\rho_{\mathrm{ye}_{\Omega}(\mathrm{x})}\rho_{\mathrm{ze}_{\Omega}(\mathrm{x})} - 2a\theta\rho_{\mathrm{ye}_{\Omega}(\mathrm{x})}\rho^{3}_{\mathrm{ze}_{\Omega}(\mathrm{x})} + a\theta^{2} - 2a\theta^{2}\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})} + a\theta^{2}\rho^{4}_{\mathrm{ze}_{\Omega}(\mathrm{x})} + 1 - \rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})} - \rho^{2}_{\mathrm{ye}_{\Omega}(\mathrm{x})}}{a(1-\rho^{2}_{\mathrm{ze}_{\Omega}(\mathrm{x})})^{2}})\Big\} = \mathbf{0}
\end{aligned}$$

(3.25)

where *a=n-q-1=degrees of freedom.* In addition, as $e_{.}(\boldsymbol{X})$ represents a regression of *Z* on ***X***, if we utilize a linear probability model to estimate the PS we have

$$\rho^2_{\mathrm{ze_{.}(x)}} = R^2_{.} \tag{3.26}$$

where $R^2_{.}$ represents the respective coefficient of determination or proportion of treatment variance explained by the predictors (e.g. Cohen, Cohen, West & Aiken, 2003). That is, because

$$R^2_{z\,by\,e(X)} = \rho^2_{\mathrm{z}\hat{\mathrm{z}}} \tag{3.27}$$

where $\hat{z}$ are the predicted values of *z* based on the regression

$$z = X\beta \tag{3.28}$$

we can reduce the correlation between the treatment and the PS to (3.26) or the correlation between the treatment and the covariates included in the PS. In other words

$$\rho^2_{\mathrm{ze_{.}(x)}} = \rho^2_{\mathrm{z}\vec{\mathrm{X}}} = R^2_{\mathrm{z\,by\,}\vec{\mathrm{X}}} \tag{3.29}$$

Using this property in conjunction with the property

$$R^2_{.} = \sum_{i=1}^{n} \beta_i \rho_{.\mathrm{x}_i} \tag{3.30}$$

we can solve for the roots in terms of $\rho_{ye(X)}$ and thus can solve for the threshold in terms of the treatment-covariate relationship (e.g. Cohen et al., 2003). Here β_i is the regression coefficient for covariate X_i obtained from regressing the treatment on the selected group of covariates and $\rho_{.\mathrm{x}_i}$ is the zero order treatment-covariate correlation. Furthermore, β_i can be obtained in a manner analogous to (3.17) through

$$\beta_i = (\frac{s_{\mathrm{z}}}{s_{\mathrm{x}_i}})(\frac{\rho_{zx_i} - \rho_{z\mathbf{x}}\rho_{\mathbf{x}x_i}}{1-\rho^2_{\mathbf{x}x_i}}) \tag{3.31}$$

In a similar manner, I provide insight into treatment effect estimators which use the PS as a weight (i.e. IPTW) in the HLM outcome model. Analogous to covariance adjustment on the PS we can estimate the treatment effect using (3.15) by dropping the *e(**X**)* term and weighting by the inverse probability of treatment. Thus *γ₁* in (3.16) can be estimated via

$$\hat{\gamma}_1 = (\frac{s_{\bar{y}}}{s_z})(\frac{r_{\bar{y}z}}{1}) \tag{3.32}$$

where $r_{..}$ represents weighted correlations. Now these weights are estimated by

$$\lambda_j * \frac{1}{P(Z_j = 1)} \tag{3.33}$$

where λ_j is obtained from (3.14) and the denominator is obtained from the estimated PS. As a result, (3.23) can be re-expressed as

$$(\rho_{yz|e_1(x)} - \theta)^2 + (\frac{1-\rho_{yz|e_1(x)}}{(n-q-1)}) = (\rho_{yz|e_2(x)} - \theta)^2 + (\frac{1-\rho_{yz|e_2(x)}}{(n-q-1)}) \tag{3.34}$$

and a threshold can be estimated by solving for $\rho_{yz|e_.(x)}$. Such derivations conceptually illustrate how regression adjustment on the PS can be reduced to a series of correlations among covariates. However, practically such approximations rely on multiple assumptions. A first assumption, is the approximation of the maximum likelihood estimator in random intercept HLMs by aggregated WLS described above. In particular, because eliminating a variable from the PS often reduces MSE by small margins, the approximation may mask such small changes. That is, excluding a single variable from the PS may only decrease the MSE by a very small amount which the WLS approximation may miss. Additionally, such derivations rely on the validity of a linear probability model in specifying the probability of receiving the treatment. That is, one could rather rely on the approximation of a nonlinear probability function such as the logit by the linear probability model. Such models tend not to be common practice as they may exceed the 0 to 1 space of probability. Further, one may again lose small yet relevant amounts of information due to the approximations. Moreover, the approximation of a logit model by a linear probability model can be especially poor when probabilities are beyond the range of 0.25 to 0.75.

Thresholds and Properties of Impact Based Construction

In identifying impact thresholds for including covariates in the PS and to estimate several properties of PS model based estimators based on such thresholds, I utilized Monte Carlo style simulation experiments. In particular, using simulated data I conducted multiple experiments to shed light on the properties and procedures of such an approach. First, I approximated thresholds for the simulated data by estimating and comparing the MSE for models that included a covariate of interest in the PS versus those that excluded it. Subsequently, I assessed how such thresholds might change when varying several parameters relevant to educational data through sensitivity analyses. Third, I assessed the power and type one error rate of constructing the PS based on the initial thresholds. Fourth, I addressed how we might consider updating the thresholds throughout the PS model building process. Fifth, I assessed the performance of utilizing proxies for the outcome to estimate the relevant relationships rather than directly using the outcome. Last, I explored what gains might be lost when a large number of potential PS variables exist and we are forced to use the thresholds in a sequential manner rather than evaluating all potential specifications. Below I describe the motivation to understand several properties of this approach.

Estimating Thresholds

To understand how PS construction based on such thresholds performs and its properties I first estimated the thresholds using simulated data. I approximated thresholds by estimating the treatment effect and its properties when the PS model contains covariates with various Γ-relationships for increasing Δ-relationships. That is, holding the Γ-relationship constant, I assessed the MSE of the corresponding treatment effect estimator for various values of the Δ-relationship. In particular, I estimated the treatment effect *m* times and then recorded the estimator's corresponding bias, variance and MSE. Subsequently, I re-specified the PS model such that it excluded the covariate of interest and estimate the treatment and corresponding properties. The threshold for a Γ-relationship is identified as the Δ-relationship in which including the covariate of interest in the PS model increases the MSE of the treatment effect estimator compared to the PS model without it. Because the Δ- and Γ- relationships are represented as partial correlations rather than zero order correlations, the dependence of thresholds on other covariates may be minimized.

Sensitivity To Parameters

Next, to understand how such thresholds may be influenced by several parameters, I assessed the sensitivity of the thresholds to several parameters that were initially fixed using simulated data. Specifically, I examined how they might be influenced by the following parameters: the between group variance, the number of groups, the probability of receiving treatment and the magnitude of the effect size. To address such variation, I explored the sensitivity of PS impact thresholds by altering the limited set of parameters above while fixing the others. In particular, I fixed the total number of individuals while requiring a minimum number individuals in each group, the variation in treatment assignment explained, the variation in outcome explained, the variation of covariate values within a group, the level one and two coefficients and the partial correlation matrices of the covariates. Through these experiments I evaluated the properties of the PS based treatment effect estimator within the HLM framework in terms of bias, variance and MSE of the estimator. Such exploration intended to shed light on variable selection in different situations.

Power and Type 1 Error

After estimating thresholds for several data schemes, I assessed the properties of the treatment effect estimator based on PS's constructed using such thresholds. The larger goal of the PS based methods is often to replicate the properties of a randomized experiment. In particular, such replication attempts to make feasible the strong ignorability assumption that randomization provides. In addition, to the strong ignorability of the treatment assignment, several other properties of randomized experiments are desirable. In particular, in replicating an experiment, such methods hope to also attain a type 1error rate consistent with the desired α-levels. In other words, given that there is no true treatment effect, we would like to incorrectly reject the corresponding null hypothesis at a level that matches the chosen type 1 error rate. In a similar manner, we would like to retain the power to detect treatment effects when they exist. For instance, in a study comparing two groups, we would like our analytic methods to match the probability in randomized experiments of rejecting the null hypothesis when the null is not true.

The effect of using an impact based approach to identify meaningfully comparable units on the type 1 error rate as well as power is of interest. As impact based construction of the PS model relies on minimizing the MSE which in turn involves the variance of the estimator, it maybe hypothesized that the type 1 error rate may be inflated. To understand how the type 1 error rate may be influenced, I conducted an additional Monte Carlo experiment comparing the type 1 error rates among the different PS construction methods. More specifically, using the impact thresholds, the stepwise procedure and the all available variables methods described previously, I estimated the type 1 error rate at 0.05 level for each method when using the aforementioned parameters and 25 variables.

Static vs. Dynamic Impact Construction

Next, I evaluated the performance of PS's constructed using the estimated thresholds in several settings. The sensitivity analyses aimed to provide insight as to the feasibility of identifying approximate thresholds for PS construction. In particular, the sensitivity analyses intend to examine whether and how the thresholds display some dependency on the parameters assessed. For instance, is it feasible for a given dataset to estimate one set of thresholds for the entire PS construction process and will such an approach fail to take into account the constantly changing values of relevant parameters? If not, the thresholds may be under or over estimated and may compromise some of the gains made by impact based construction. That is, for a given dataset it may be most beneficial to adjust one's thresholds to the current PS iteration in order to account for any changes in parameters and thresholds. In other words, the thresholds for including a variable depend on which variables you have already included through the Γ and Δ quantities but also depend on how previously included variables changed other relevant parameters. This gives rise to a certain static versus dynamic impact based construction of the PS. Put differently, using a static approach one might identify thresholds at the beginning of the PS construction phase and utilize those thresholds for each iteration in the PS construction process. Alternatively, a more dynamic approach might be to re-identify such thresholds at each iteration. In other words, after including a covariate in the PS, we might not only update the Γ- and Δ-relationships but also update the other parameters discussed so that our thresholds consider both the change in the partial correlations as well as change in the other parameters. In this way, one might call this a more dynamic approach in that it updates all parameters involved.

In addition, though optimally one would consider and assess all potential sets of covariates, it is often more practical to construct a PS model in a step by step fashion. As a result, even when taking into account the stable (observed) quantities in one's data such as sample size, the PS model building procedure sequentially modifies multiple parameters, e.g. amount of variance explained, continuously requiring adjustments to such thresholds. As a result, in order to effectively and consistently decrease the MSE of the treatment effect estimator using impact based construction one may need to trade static thresholds for more dynamic thresholds. In other words, if the thresholds illustrate some dependency on the sensitivity parameters, an alternative approach would be to adjust thresholds to align most closely with each phase of construction. Identifying such thresholds would require constant assessment of the sensitivity parameters. Such an approach could be simplified by directly comparing the MSE of each estimator, thereby implicitly using the thresholds. That is, we would include a variable in the PS only if it directly decreases the MSE. Formally, covariate X_n is included in the PS only if the

$$MSE(\hat{\gamma}_{e_{\omega}(\vec{X})}) \geq MSE(\hat{\gamma}_{e_{\Omega}(\vec{X},X_n)}) \tag{3.35}$$

However, such an approach assumes a priori knowledge of the true treatment effect. As a result, a more tractable approach may be to identify thresholds for a range of parameters. To assess and compare the performance of both the dynamic and static impact methods in conjunction with the all and step methods, I conducted an additional simulation. That is, in the static approach I re-estimated only the Γ-and Δ- relationships after a new covariate is added to the PS and did not re-estimate the other relevant parameters but rather rely on the initial and fixed estimates and their corresponding thresholds. In contrast, the dynamic approach re-estimated each parameter and the corresponding thresholds. To understand how the dynamic approach sustains the intended reduction in MSE while the static approach deflates the intended reduction in MSE, this simulation was done with a large number of available potential covariates. In other words, it is likely that the benefits from continuously updating the parameter estimates and their respective thresholds may only be realized when there are a number of adjustments.

Use of Outcome Proxies

Above I discussed several salient features of educational data that make it plausible to use proxies for the outcome to estimate relevant relationships. Specifically, I offered three different potential approaches: using prior ability as a proxy for the outcome when it represents a similar ability meausre, using prior waves of data as proxies for the outcome and covariates, and using a subsample for cross validation. Each approach has some desirable properties as well as some sources of uncertainty. In particular, the proxy approach relies heavily on the availability of a measure known to be highly correlated with the outcome and likely correlate with other covariates in a similar manner. However, when available, this approach allows one to construct the model using the actual data collected and allows one to retain the full sample size. Similarly, using prior waves of data to construct the PS allows one to retain the full sample for the outcome analysis but admits uncertainty as the data from similar studies or prior waves may not capture the idiosyncratic features as the current data. In contrast, the cross-validation approach requires one to effectively give up a portion of the collected sample size. In doing so one adds uncertainty in terms of a reduced sample size, but one also eliminates uncertainty as we are able to directly use the outcome rather than a proxy.

To attend to these tradeoffs, I conducted a simulation assessing the proxy approach with the cross-validation approach using each of the construction methods and PS uses. Specifically, I assessed the MSE of the treatment effect estimator when using a proxy outcome, e.g. pretest, with a correlation of 0.70 with the outcome versus a cross-validation approach using a random 50% sub-sample. Further, each experiment used a group size of 500 and estimated a MSE using 500 simulations.

Stepwise

A second issue that arises when practically implementing the impact based approach is model selection. Though not specific to the impact based construction method, developing a model that measures some criterion, be it R^2, AIC, or in the impact case the MSE of the treatment effect estimator involves multiple decisions, most of which usually hinge on prior decisions. Model selection approaches should assess the appropriate criterion for all possible models and select the model that optimizes the criterion. However, with large sets of observations and variables, such an approach is computationally difficult. As a result, such approaches often rely on suboptimal procedures such as a step by step approach. Although impact based construction addresses this through the way it characterizes the relevant relationships, i.e. conditional relationships, it is not fully immune to pitfalls faced by other stepwise procedures. In particular, one can envision a scenario where an impact constructed PS model depends on the order in which one considers potential covariates. For instance, suppose we have only two covariates and that the first, X_1, exceeds the appropriate impact threshold and thus should be included in the PS model. However, suppose we decide to first consider covariate X_2 and it as well exceeds the appropriate threshold. After inclusion of X_2 in the PS model, we re-assess the X_1 covariate and now find that it no longer exceeds its threshold. In this example, the construction of the PS depends on the order in which we consider potential covariates. In order to resolve this, the approach most faithful to the impact based method and its corresponding minimization of MSE would be to consider all possible combinations. For example, one would consider sets of covariates. That is, because the Γ- and Δ- relationships are conditional relationships, one can additionally consider adding a group of covariates all at once. In such cases we simply define the Γ- and Δ-relationships as the collective correlation coefficients which are conditional upon those covariates already in the PS model. As a result, similar to other model building procedures, we could consider all potential combinations of covariates. However, such an approach is practically infeasible in all but the largest studies and even then remains computationally expensive with even a moderate number of measured variables. Consequently, for larger datasets the impact based construction method can be practically implemented using procedures similar to a stepwise approach. Whereas stepwise selects models based on AIC or p-values, the impact based criteria selects models that improve the treatment effect estimator in terms of MSE. In other words, the impact based approach swaps one criteria for another but does so in a manner that places focus on the quantity of ultimate interest. To examine the effect of sequentially adding covariates to the PS based on thresholds, I conducted another simulation.

Simulation Design

I simulated data to embody several typical instances in educational data. In particular, when such literature was available, I based parameter values on educational literature which has demonstrated common values for the parameters I considered. Each of the simulation experiments employed similar data generating processes. I generated the individual characteristics, ***X****, (e.g. X_1, X_2, X_3) and group characteristics, ***W****, (e.g. W_1, W_2, W_3) from two multivariate uniform distributions within the range [-5,5]. Subsequently, for a given experiment and threshold, I specified the desired partial correlation matrices $\sum_{x|\mathbf{X}}$ and $\sum_{w|\mathbf{W}}$:

$$\sum_{x|\mathbf{X}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \qquad \sum_{w|\mathbf{W}} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \tag{3.36}$$

where the elements of $\sum_{x|\mathbf{X}}$ and $\sum_{w|\mathbf{W}}$ represent the partial correlation controlling for other variables respectively. Using

$$r_{xy} = r_{xy|\mathbf{X}}\sqrt{(1-r_{x\mathbf{X}}^2)(1-r_{y\mathbf{X}}^2)} + r_{x\mathbf{X}}r_{y\mathbf{X}} \tag{3.37}$$

to specify the zero order correlations, we have zero order correlation matrices $\sum_x$ and $\sum_w$:

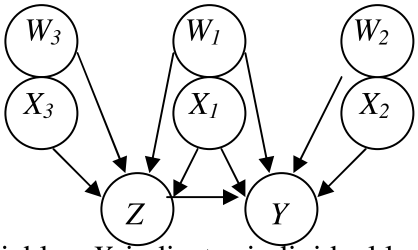
$$\sum_{\text{x}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \qquad \sum_{\text{w}} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \tag{3.38}$$

Using ***X****, ***W****, $\sum_{\text{x}}$ and $\sum_{\text{w}}$ and Cholesky decomposition I generated ***X*** and ***W*** such that they conform to $\sum_{\text{x}}$, $\sum_{\text{w}}$, $\sum_{\text{x}|\mathbf{X}}$ and $\sum_{\text{w}|\mathbf{W}}$.

The treatment was designed such that it represents the realization of a dichotomous variable given group characteristics. For all experiments, the true treatment assignment mechanism will follow the generalized linear model:

$$logit(P(Z=1)) = \gamma_0 + \sum_{j=1}^{q_w} \gamma_j W_j \tag{3.39}$$

for q_w group level covariates. As depicted by Figure (3.40), three types of variables were considered with various magnitudes: true confounders (e.g. W_1), those only related to the outcome (e.g. W_2) and those only related to the treatment (e.g. W_3). Though some variables are not directly related to the treatment assignment (e.g. W_2) all variables possess some empirically non-zero relationship due to chance.



Figure(3.40): Relation of variables: *X*. indicates individual level characteristics, *W*. indicates group level characteristics, Z indicates the treatment and *Y* is the outcome

The true outcome model for each experiment was a hierarchical linear model with a random intercept

$$
\begin{aligned}
&Level\,1: Y_{ij} = \beta_0 + \sum_{p=1}^{q_x} \beta_i X_{pij} + \varepsilon_{ij} \\
&Level\,2: \beta_0 = \gamma_0 + \gamma_1 Z_j + (\sum_{j=1}^{q_w} \gamma_j W_j) + u_{0j}
\end{aligned}
\qquad (3.41)
$$

where Z is the treatment effect, $e \sim N(0, \sigma^2_{outcome})$, $u_0 \sim N(0,\tau_{outcome})$ and excluded are those variables unrelated to the outcome depicted in Figure (3.40) (e.g. X_3 and W_3).

For each experiment, I estimated the treatment effect for each data set and each PS method (i.e. stratification, matching, IPTW, covariance adjustment) using a HLM (Hong & Raudenbush, 2006). Specifically, when stratifying or matching, I modeled the outcome as a HLM as follows:

$$
\begin{aligned}
&Level\ 1: Y_{ij} = \beta_{0j} + \sum_{p=1}^{q_x} \beta_{pj} X_{pij} + \varepsilon_{ij} \\
&Level\,2: \beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + (\sum_{i=1}^{s} \gamma_{0s} S_{sj}) + u_{0j}
\end{aligned}
\qquad (3.42)
$$

where S represents either the number of strata or matched groups minus one for the subclassification or matching approach, respectively. When considering covariance adjustment on the PS, I use the HLM estimator

$$
\begin{aligned}
&Level\,1: Y_{ij} = \beta_0 + \sum_{p=1}^{q_x} \beta_i X_{pij} + \varepsilon_{ij} \\
&Level\,2: \beta_0 = \gamma_0 + \gamma_1 Z_j + \gamma_2 e(\mathbf{W}) + u_{0j}
\end{aligned}
\qquad (3.43)
$$

where $e(\mathbf{W})$ is the estimated PS for each group. When considering IPTW, I use (3.43) without the PS but weight each group appropriately. Using the results of the Monte Carlo simulation experiment, I estimated the bias and mean-squared error (MSE) of each approach using:

$$
\widehat{Bias}(\hat{\boldsymbol{\theta}}) = \frac{1}{M} \sum_{i=1}^{M} \hat{\gamma}_i - \gamma_{true} \qquad (3.44)
$$

and

$$\widehat{MSE}(\hat{\boldsymbol{\theta}}) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\gamma}_i - \gamma_{true})^2 \tag{3.45}$$

where *M* represents the number of simulated data sets and the variance of the estimator can be estimated using (3.13).

Results

Thresholds

I present the results in six parts. First, I examined the estimated thresholds along with their trends and differences among the different uses of the PS. Second, I examined the sensitivity of the estimated thresholds to the variation of those parameters listed above. Third, I present the estimated power and type one error rate for the simulated data. Fourth, I explore the benefits and drawbacks of using a static versus a dynamic approach to PS model building based on the estimated thresholds. Fifth, I discuss the performance of the estimator when estimating variable relationships via a proxy variable or cross validation. Sixth, I present the results of using a sequential approach to constructing PS's based on the estimated thresholds.

To make simulations tractable, I initially assumed the values for those parameters that were allowed to vary. I subsequently discuss how their values varied and what their variation may imply in the sensitivity analyses. First, the probability of receiving treatment, *p*, was set to 0.5 to reflect a 50% chance of receiving the treatment. Second, the correlation coefficient between the logit of the treatment and covariates in the PS model save the one of interest, Γ^*, was set to 0.3 and the correlation coefficient between the outcome and covariates in the PS model save the one of interest, Δ^*, was set to 0.7. Taken together with their subsequent ranges in the sensitivity analyses, such values represent treatments and outcomes where there is little understanding of their variation to where there is a great deal of explanation as to why they vary. Next, the correlation of the covariate of interest and the logit of the treatment, Γ, was given an initial value 0.3. Such a value tends to represent a moderate correlation in the social sciences and taken together with the sensitivity analyses the range represents small, moderate and large relationships in the educational data (Cohen & Cohen, 1988). Similarly, the treatment has an initial effect size, δ, of 0.3 representing moderate effect whereas the sensitivity analyses examine a small and large effect. The intra-class correlation (*ICC*) was set to 0.2 to reflect the common variance partition found in educational achievement data (Coe & Makoto, 2009). I set the group sample size, n_j, to 100. Again such a value represents the middle ground between the values used in the sensitivity analyses. Collectively, such group sizes were chosen to represent a lower bound for effective use of multilevel models and a practical upper bound for educational data (Maas & Hox, 2005; Moinedden, Matheson & Glazier, 2007).

Figure (3.46) displays the relationship between the MSE and the Δ-relationship for a covariate with a Γ-relationship of 0.3. Specifically, the y-axis displays the difference in MSE (i.e. the MSE of the estimator based on the larger PS, Ω, minus the MSE of the estimator based on the smaller PS, ω) and the x-axis represents the Δ-relationship of the covariate of interest (partial correlation between the covariate and the outcome). The threshold is identified as the point in which the curve crosses the horizontal line of zero, indicating that the MSE(Ω) is equal to the MSE(ω). In other words, when a curve is below the x-axis the covariate should be included in the PS model as it will decrease the MSE of the corresponding treatment effect estimator. Conversely, when the curve is positive the covariate should not be added to the PS model as its inclusion will tend to increase the MSE.

Figure(3.46): Change in MSE of the HLM estimator as a function of the covariate's Δ-relationship. Change in MSE represents the MSE of the HLM estimator based on the large PS model minus the MSE of the HLM estimator based on the small PS model.

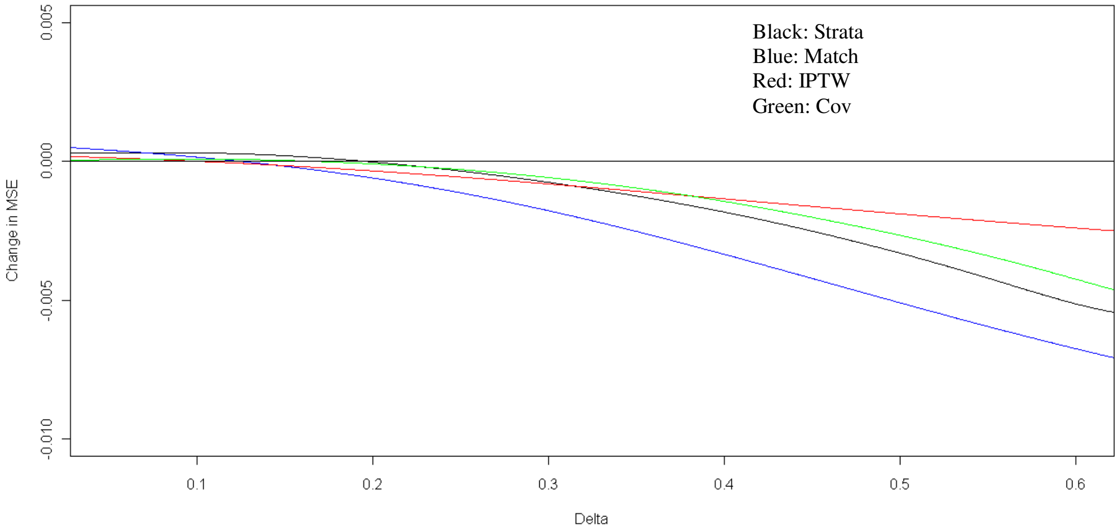


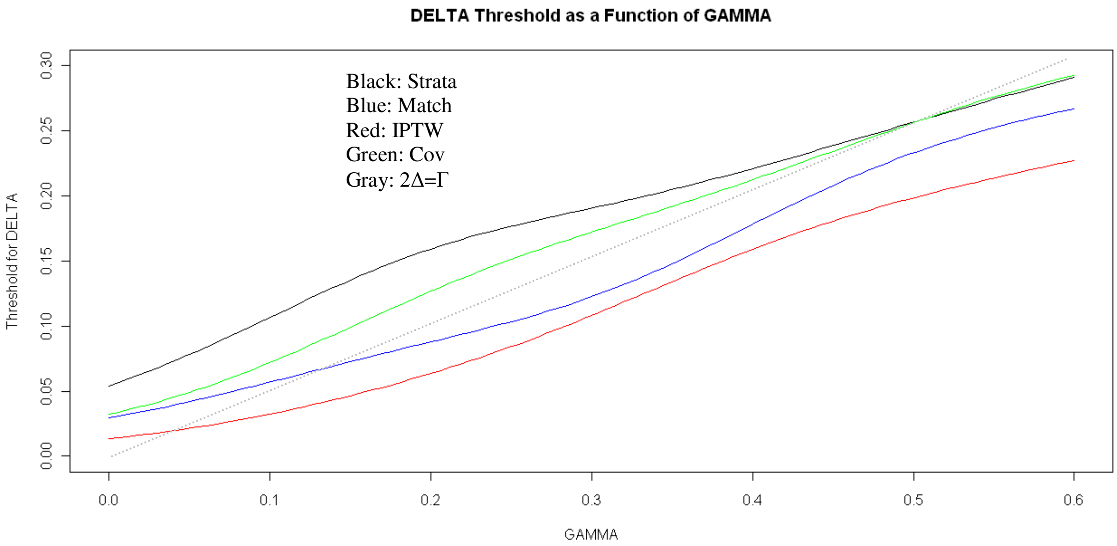
Figure (3.46) and similar figures suggest several interesting properties in this data. First, excluding a single covariate from the PS model potentially has a considerable amount of influence on the MSE of the corresponding treatment effect estimator. However, the magnitude of such influence heavily depends on the covariate's Δ-relationship.

Similarly, from figure (3.46) it is also evident that although there is a threshold, the decision to include or exclude a single variable will have a small effect on the MSE of the estimator unless that variable has a strong relationship with the outcome. However, when we compare this added MSE with the magnitude of the MSE for the correct PS model its addition is considerable. For example, given the above parameters including a true treatment effect size of 0.30, constructing the PS using a single additional variable which has a Γ-relationship of 0.30 and a Δ-relationship of 0.1 and subclassifying on it in an HLM outcome model would add roughly $\sqrt{0.0015}$ or 0.039 to the average error. In other words, in estimating a treatment effect of 0.30 the inclusion of a single additional covariate that falls short of the threshold in the PS would increase the error by roughly 10% of the treatment effect. Put on a different scale, adding a covariate to the PS model when its Δ-relationship does not meet the minimum threshold value for its corresponding Γ-relationship would increase the overall MSE of the treatment effect estimator by over 26% ($\sqrt{0.0025}$ or 0.05 versus $\sqrt{0.0025+0.0015}$ or 0.063). Though this additional error is relatively small, when considering the inclusion of many covariates into the PS, the contribution to MSE of covariates that do not meet their respective thresholds is potentially summative. That is, because the Δ- and Γ-relationships are specified as conditional relationships each covariate's contribution to the overall MSE may be somewhat unique. In other words, the change in MSE represents the average additional change in MSE for the inclusion of a single variable with specified Δ- and Γ-relationships (or set of variables with the said Δ- and Γ-relationships). As a result, adding several variables to the PS which do not attain the threshold can exceedingly detract from the quality of the estimator. For instance, assuming the error is additive, adding ten variables which fail to attain their respective thresholds to the PS will increase the average error of the estimator by more than $\sqrt{0.0015*10}$ or 0.12. In the current context, assuming those value of Δ and Γ above, the average error would be $\sqrt{0.0025+10*0.0015}$ or 0.13. Such additional error represents more than a 200% increase in the average error of the estimator. More generally, the magnitude of such additional error represents more than 30% of the actual treatment effect, 0.30. In other words, for this simulated data if the structural form of the outcome model were correct and we used an unbiased estimator of the treatment effect, our construction of the PS in a manner that neglects such threshold considerations might, on average, cause us to misestimate the treatment effect by approximately 0.13 if there were 10 variables in the PS that did not meet their respective thresholds.

Next, from Figure (3.46) we tend to see an ordering of the thresholds corresponding to the different PS uses. In particular, the PS use with the highest threshold tends to be subclassification on the PS. In other words, including covariates in the PS model when one intends to use subclassification on the PS in an outcome model requires higher Δ-relationships relative to the other uses (e.g. matching, IPTW, covariance adjustment). As subclassification tends to be a coarse adjustment, the MSE is decreased only when the variable is at least moderately related to the outcome. Corresponding to the high threshold for subclassification is a diminished change in MSE as the Δ-relationship increases relative to the other PS uses. Put differently, using subclassification on the PS one tends to incur a smaller, gradual increase in MSE by excluding covariates with moderate to high Δ-relationships relative to the other uses. This is evident in Figure (3.46) as subclassification curve tends to decrease at a rate slower than the other uses. Such results also suggest that subclassification on the PS may provide the most protection against unmeasured confounding variables. In contrast, using the PS as IPTW in an outcome model tends to have lower thresholds but increased MSE at lower Δ-relationships levels. Put another way, if an analyst intends to use IPTW on the PS in the outcome model, he/she should include covariates in the PS model with smaller Δ-relationships than if he/she were to use subclassification on the PS. In addition, if using the PS for IPTW the exclusion of covariates with moderate to strong relationships with the outcome will tend to result in a MSE that is higher relative to the corresponding MSE based on subclassification. In other words, IPTW tends to be a more fine-tuned adjustment as compared with subclassification.

Though Figure (3.46) illustrates the thresholds for several different uses of the PS, it does not show how these thresholds change for different parameters. A particularly relevant parameter when identifying such thresholds is the magnitude of the Γ-relationship. As the MSE of the treatment effect estimator is influenced by both the Δ- and Γ-relationships, minimizing the MSE requires the exclusion of those covariates whose decrease in bias is exceeded by its increase in variance. Figure (3.47) displays how the threshold changes as a function of Γ-relationship.

Figure(3.47): Δ-relationship thresholds as a function of a covariate's Γ-relationship.



Similar, to previous thresholds, we see that subclassification on the PS tends to have the highest threshold for most of the Γ-relationships. In other words, for a covariate with a given Γ-relationship, the magnitude of the Δ-relationship has to be higher for the inclusion of a given covariate to decrease corresponding MSE when using subclassification compared to the other uses. Additionally, the thresholds tend to increase at a similar rate and, for the most part, preserve their order. That is, subclassification tends to have the highest threshold, followed by covariance adjustment, followed by matching, and finally followed by IPTW. These relationships suggest that although a covariate may explain portions of the treatment assignment mechanism, its value in estimating the treatment effect using a PS based estimator is modified by its relationship to the outcome. As a result, covariates strongly related to the treatment assignment mechanism may actually decrease the quality of the estimator in terms of MSE if they do not have corresponding relationships similar in magnitude with the outcome. Though the thresholds vary, a rough rule of thumb for assessing the value of a covariate in similar data is that the Δ-relationship should be at least half the size of the Γ-relationship in order for its inclusion in the PS to decrease the MSE of the corresponding treatment effect estimator. This rough rule could be further supplemented by recognizing that when using subclassification or covariance adjustment the Δ-relationships should, for the most part, be slightly above half the magnitude of the Γ-relationships and when using matching or IPTW the Δ-relationship should be slightly below half the magnitude of the Γ-relationship. Table (3.48) provides a summary of the thresholds for four different uses of the PS. Table (3.48) combines several values of the Γ-relationship along the top with four different uses of the PS along the left side and lists the corresponding Δ-relationship thresholds inside the table. In other words, values inside the table represent the minimum Δ-relationships covariates must have for a given Γ-relationship and PS use in order for them to reduce the MSE of the corresponding treatment effect estimator.

Table(3.48): Impact construction thresholds for default simulation

	Γ=0.05	0.10	0.15	0.20	0.25	0.30	0.40	0.50	0.60
Subclassification	0.06	0.10	0.13	0.16	0.17	0.19	0.22	0.26	0.28
Covariance	0.05	0.07	0.10	0.12	0.15	0.17	0.22	0.26	0.28
Matching	0.04	0.06	0.07	0.09	0.10	0.12	0.17	0.23	0.27
IPTW	0.02	0.03	0.04	0.07	0.08	0.11	0.16	0.20	0.23

Comparing Approaches

To attend to how this approach compares to other standard approaches, I first examined conceptually how the impact based construction method measures the treatment effect estimator. Second, for each data set and use of the PS, I compare the bias, variance and MSE of models constructed by impacts with those constructed by standard PS model building procedures by contrasting their estimated densities. In particular, I consider two alternative PS construction methods: a stepwise AIC approach (e.g. Hastie & Pregibon, 1992) and an including all available variable approach.

Figures (3.49) to (3.56) conceptually contrast the different PS model construction approaches in terms of MSE. Specifically, each figure displays the MSE of the HLM treatment effect estimator based on each of the three PS construction methods and Γ-relationship as a function of the Δ-relationship. For instance, for covariates with a Γ-relationship of 0.10 Figure (3.49) displays the average MSE of the treatment effect estimator as a function of the stepwise approach (blue), the all approach (black) and the impact approach (red). In these figures, the y-axis represents the average MSE of the HLM treatment effect estimator and the x-axis represents the Δ-relationship for a covariate we may include in the PS. In Figure (3.49), the blue line represents the MSE for a PS model constructed using the stepwise approach. I note that for Figure (3.49) the all approach (black line) is plotted on top of the blue line until the Δ-relationship reaches

0.10 and plotted on top of the red line thereafter. Specifically, the stepwise PS model tends to exclude the covariate of interest, regardless of its Δ-relationship. As a result, we see that it lowers MSE only when the Δ-relationship is less than about 0.10. The black line in Figure (3.49) represents the MSE for a PS model that includes all available variables regardless of their Δ-relationships. This approach tends to reduce MSE only if the covariate has a Δ-relationship greater than 0.10. Finally, the red line represents the MSE for a PS model constructed using the impact approach. Specifically, this approach includes only those covariates whose Δ-relationship exceeds the threshold corresponding to its Γ-relationship. Evident from the Figure (3.49), for a Γ-relationship of 0.10 the impact approach is similar to the stepwise approach until it reaches the Δ threshold and then is similar to the all approach. In particular, when focusing on the inclusion/exclusion of a single variable with a Γ-relationship of 0.10, the stepwise approach will almost always exclude the potential covariate regardless of it Δ-relationship. The effect of excluding the potential covariate is highly dependent on the Δ-relationship. That is, when the Δ-relationship is small it can actually be beneficial to exclude the covariate, whereas when the Δ-relationship is high it can be costly to exclude the covariate. Alternatively, the all approach incurs a MSE higher than stepwise for small values of the Δ-relationship but clearly dominates the stepwise approach when the Δ-relationship is high. Figure (3.49) clearly depicts the bias-variance tradeoff that we are trying to balance. Specifically, the impact based approach conceptually attempts to move between these approaches in a manner that always minimizes the MSE. That is, the impact based approach conceptually chooses the estimator with the minimum MSE. To see this in Figure (3.49), I note that the red line representing the impact approach lies directly on the blue line representing stepwise until about a Δ-relationship of 0.10. Thereafter, the red impact line switches and lies directly on the black line representing the all approach. For example, for any Δ-relationship greater than 0.10, the impact approach will make inclusion decisions identical to the all approach. Similarly, though not generally true, in Figure (3.49) the impact approach tends to make decisions similar to the stepwise approach when the Δ-relationship is less than 0.10. Subsequent figures with Γ-relationships of 0.10 illustrate similar concepts but adjust the point at which the impact approach switches from the stepwise approach to the all approach based on the method by which we use the PS in the outcome model.

Figures (3.49) to (3.52): MSE of the treatment effect estimator as a function of the outcome-covariate relationship for each construction method for a covariate with a treatment-covariate relationship of 0.10.

Figure(3.49)

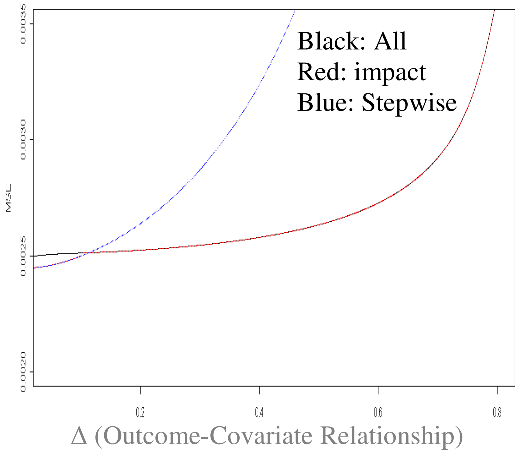
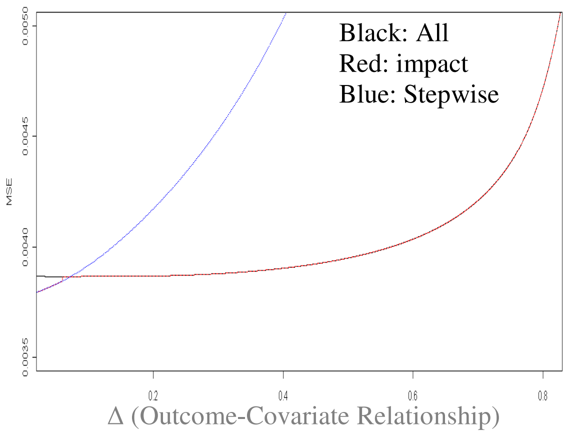
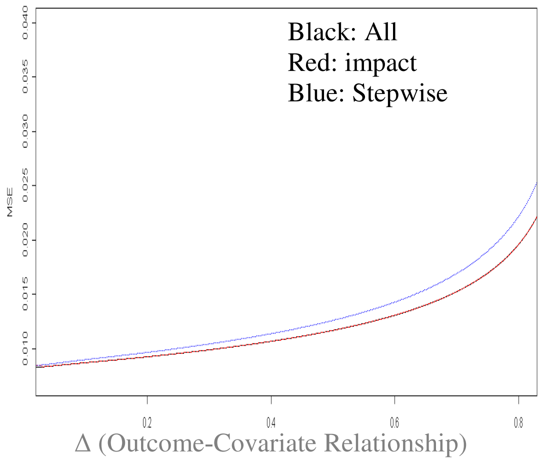


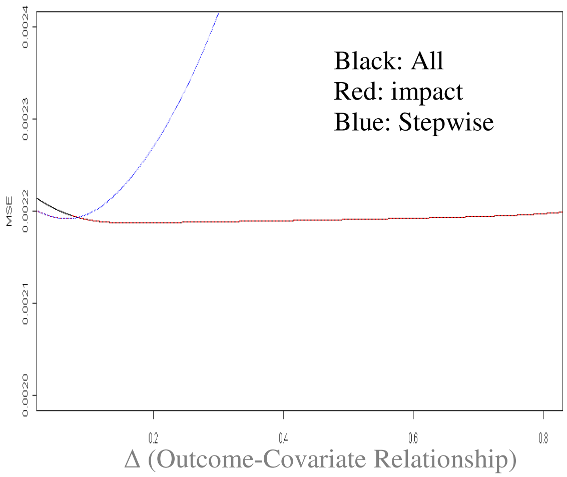
Figure (3.50)



Figure(3.51)

Figure (3.52)





In a similar manner, Figures (3.53) to (3.56) illustrate the tradeoff that minimizes the MSE of the treatment effect estimator for Γ-relationships of 0.30. For example, in Figure (3.53) the black line represents the MSE of the all approach. As the stepwise approach generally includes a covariate in the PS model if it has a partial correlation with the outcome of 0.30, the line depicting the stepwise approach lies on the all approach black line. In other words both the stepwise and all approaches will include the covariate regardless of the Δ-relationship. In contrast the impact approach (red) only includes covariates that meet or exceed the threshold corresponding to their Γ-relationships. As a result, for covariates with a Γ-relationship of 0.30 it excludes covariates whose Δ-relationship is less than 0.19. Thus, in this example, the impact approach deviates from both the stepwise and all approach when the Δ-relationship is less than 0.19 and then merges with the other two approaches thereafter. Such a hybrid approach minimizes MSE for a given set of potential covariates to be included in the PS model.

In comparing the performance of impact based approach with the stepwise and all approaches, we see that the relative gain depends to a large extent on the Γ-relationship and to some extent on the method by which we use the PS in the outcome model. When the Γ-relationship is a 0.10 and we are subclassifying on the PS, we see about a 4% decrease in MSE when comparing the impact approach with the all approach for Δ-relationships less than 0.10. When contrasting the impact based approach with the stepwise approach, we see the impact based approach offering increasingly large decreases in MSE after the threshold. In other words, though the impact based approach provides a considerable reduction in MSE over the all approach, it provides a much larger advantage when compared to the stepwise approach. When we increase the Γ-relationship to 0.30 (Figure (3.53) we tend to see larger reductions in MSE. As the step and all approaches are similar in this context, I only contrast the impact with the other two. Specifically, though the reduction depends on the magnitude of the Δ-relationship, we see about a 10-15% reduction in MSE for covariates whose Δ-relationship does not meet the threshold.

Contrasting such gains across uses of the PS we see similar, but varying, levels of reduction in MSE. For instance, with matching, IPTW and covariance adjustment, the reduction in MSE from using the impact based approach over the others is slightly smaller than that of subclassification and again depends largely on the magnitude of the Γ-relationship.

Figures (3.53) to (3.56): MSE of the treatment effect estimator as a function of the outcome-covariate relationship for each construction method for a covariate with a treatment-covariate relationship of 0.30.

Figure(3.53)

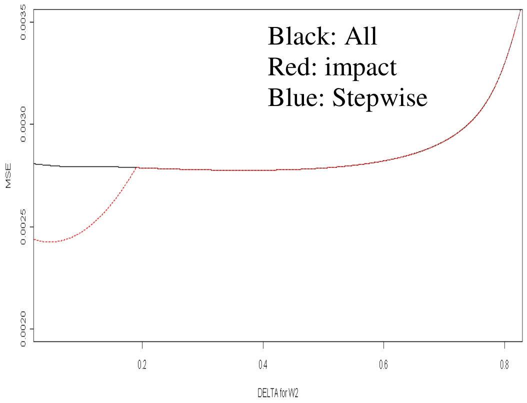
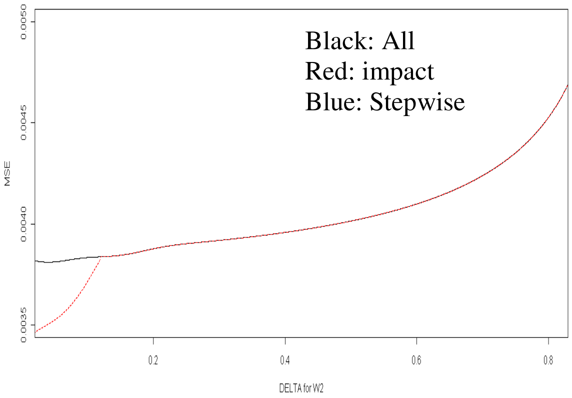


Figure (3.54)



Figure(3.55)

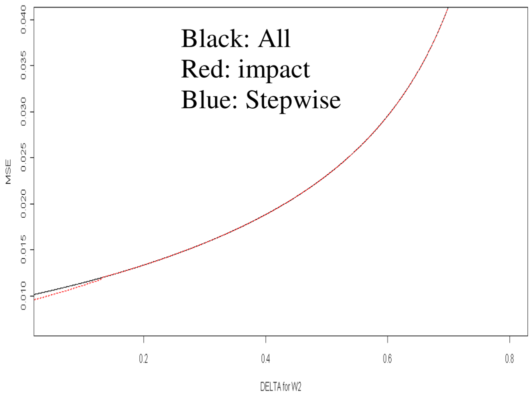
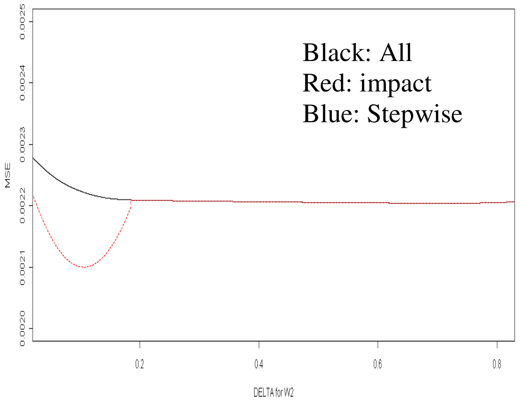
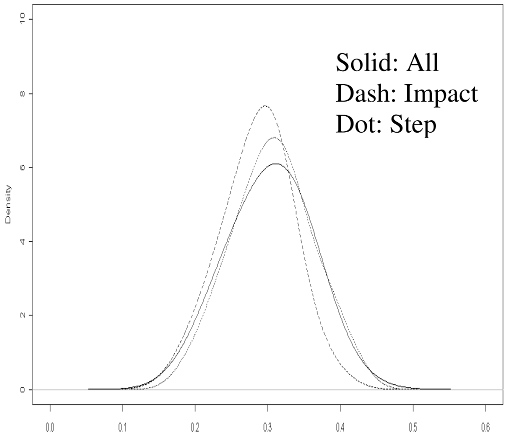


Figure (3.56)

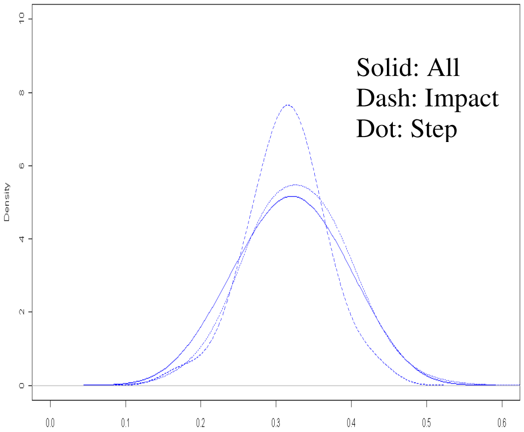


In a similar manner, I next directly compared the bias, variance and MSE of the HLM estimators based on the PS constructed by impacts with those constructed by standard PS model building procedures by contrasting their estimated densities. Figures (3.59) to (3.58) display the HLM estimator densities based on the PS construction type for each of the PS uses. For all PS uses, we see several properties of the estimators based on how the corresponding PS was formed. The estimator based on constructing the PS using the impact based approach has the most density in the neighborhood of the treatment effect, δ =0.3, in that it has least dispersion around its mode. In contrast, the all available variables approach tends to be the most dispersed regardless of the PS use and the step approach falls in between the other two.

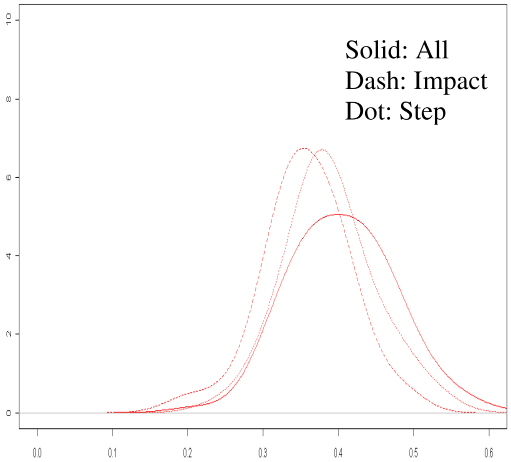
Figure(3.57): Densities for IPTW



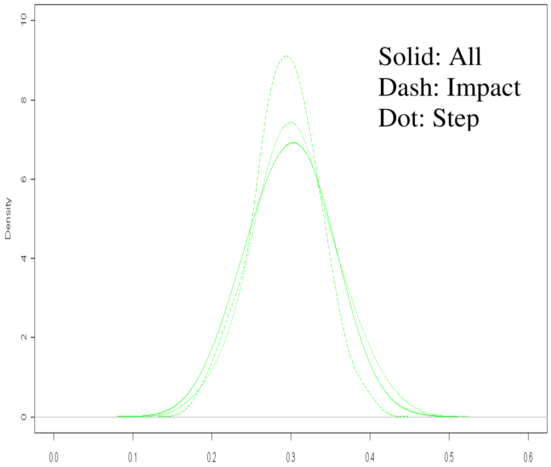
Figure(3.58): Densities for Covariance Adjustment



Figure(3.59): Densities for Subclassification



Figure(3.60): Densities for Matching



In addition, Table (3.61) estimates the bias, variance and MSE of the each PS use and construction method. Specifically, I considered the properties of the HLM random intercept estimator using covariance adjustment on individual level variables and the four different PS adjustments. Though the simulations indicate some finite sample bias when using the IPTW use, all estimators appear to be virtually unbiased. As a result, the main contribution to the MSE is the variance of the estimator. Further in contrasting the MSE of each estimator, Table (3.61) suggests that although the magnitude depends on the PS use, the impact based construction method consistently provides an estimator with lower MSE.

Table(3.61): Bias, variance and MSE of the treatment effect estimators by PS use and PS construction method

		Strata			IPTW			Cov			Match	
	All	Step	Impact	All	Step	Impact	All	Step	Impact	All	Step	Impact
Bias	0.0024	0.0038	0.0050	0.1053	0.0812	0.0658	0.0025	0.0049	0.0065	0.0200	0.0284	0.0160
Var	0.0025	0.0023	0.0021	0.0050	0.0027	0.0033	0.0020	0.0025	0.0019	0.0035	0.0037	0.0027
MSE	0.0025	0.0024	0.0022	0.0161	0.0093	0.0076	0.0020	0.0025	0.0019	0.0039	0.0045	0.0029

Type 1 Error

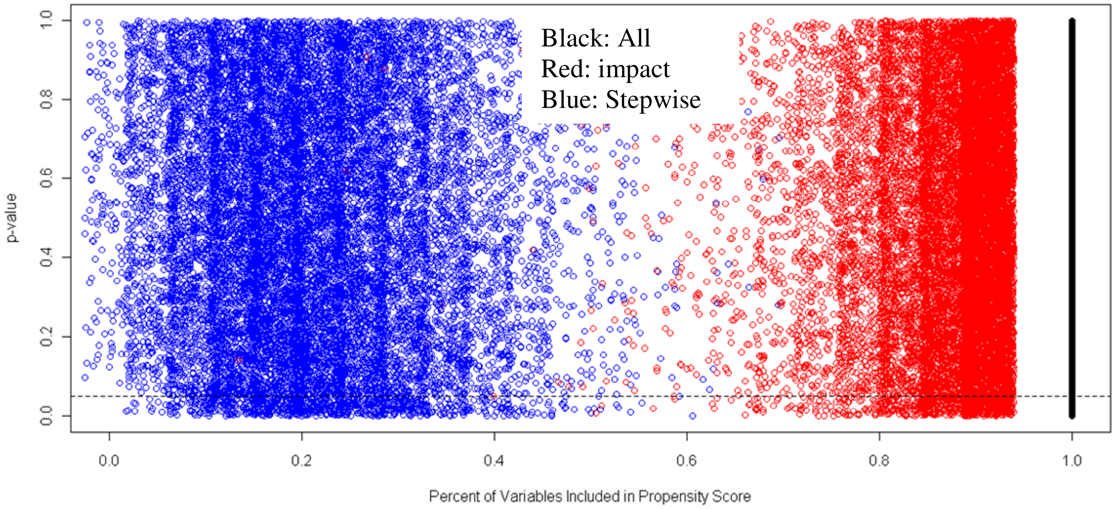
I next turn to the power and type one error rate of the impact method. The results for the type 1 error rates for each combination of PS construction method and PS use in the outcome model are displayed in Table (3.62). In general, we see that the type 1 error rates tend to be close to the target α-level of 0.05. However, we do see slightly inflated and deflated type 1 error rates depending on both the construction method as well as the end use of the PS in the outcome model. More specifically, the stepwise PS construction method tends to have a slightly inflated type 1 error rate indicating that it rejects the null hypothesis more than the nominal 5% level. Though the type 1 error rate varies by how one uses the PS in the outcome model, we consistently see that using the stepwise construction method inflates the type 1 error by slightly less than 1%. The impact construction method tends to have a slightly less consistent trend across PS uses. In particular, when using the impact construction method combined with subclassification or covariance adjustment, the type 1 error rate is slightly deflated. In contrast, using the impact method with the IPTW tends to slightly inflate the error whereas using it with matching tends to be consistent with the nominal α-level.

Table(3.62): Type 1 Error Rates (α=0.05)

	Strata	Match	IPTW	Cov	Average
All	0.044	0.046	0.045	0.043	0.045
Impact	0.049	0.050	0.058	0.049	0.051
Stepwise	0.059	0.058	0.055	0.059	0.058

It is also of interest to examine how such type 1 errors are distributed among variable selection. Specifically of interest is whether eliminating an increasing percentage of the potential PS construction variables inflates or deflates the type 1 error rate. Figure (3.63) graphically summarizes the type 1 error rates for all construction methods by plotting the treatment's p-value as a function of the percent of potential variables included in the PS. That is the blue dots indicate the p-values relative to the percent of all the available covariates the stepwise method selected to construct the PS; the red dots indicate the same but for the impact method; and finally the black dots indicate the p-values when including all the available variables in the PS. Generally we see type 1 error rates similar to the nominal level of 0.05. In addition, though obviously dependent on the data, the nature of the plot allows us to gauge how many variables each method tends to include in the PS. In particular, I note that the stepwise approach to constructing the PS values parsimony as it tends to use about 10-30% of the available variables. Substantially higher is the impact approach as it tended to use between 80-90% of the available variables. Despite the difference in the number of variables used, we tended to see a close alignment of both point estimates and inferences resulting from the all and impact methods. Such alignment presents an interesting result. Specifically, the alignment of inferences resulting from constructing the PS using all available variables and just the impact variables suggests that the impact relationships may adequately capture the relevant imbalances between groups. In other words, one could construct the PS using all available variables to provide maximal protection against biased estimates, however, one does so at a cost and that cost is a loss in efficiency. In contrast, one could capture virtually the same protection with a reduced cost in terms of efficiency using impact based construction. As a result, we can decrease the MSE of our estimates without a significant associated inflation of the type 1 error.

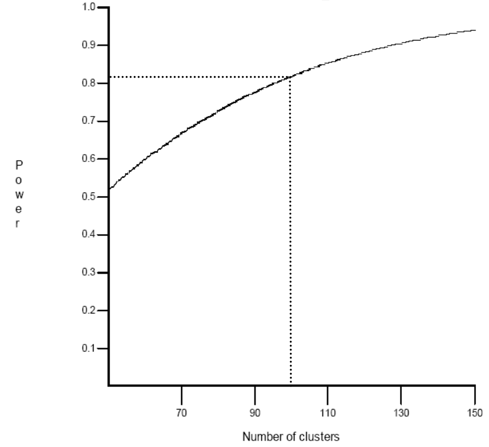
Figure(3.63): Type 1 error rates for stepwise, impact and all based PS construction methods



Power

In a similar manner, it is of interest whether the impact approach compromises the power in detecting a treatment effect. To identify the theoretical power of a cluster-randomized experiment that the current PS and HLM methods try to mimic, I estimated the approximate power of a cluster-randomized trial with similar parameters. Using Optimal Design (Spybrook, Raudenbush, Liu & Congdon, 2006), the approximate power of such a cluster randomized trial is slightly above 0.80 (Figure (3.64)).

Figure(3.64): Power as a Function of the Number of Clusters for the Hypothetical Cluster-Randomized Experiment the Current Study Attempts to Replicate



In other words, the probability of rejecting the null hypothesis of no treatment effect when in fact there is a treatment effect of 0.30 is approximately 0.8. Similar to the simulation assessing the type 1 error rate, I estimated the power of each PS construction method in conjunction with each PS use in the HLM using simulations. Table (3.65) summarizes the power for each combination. Similar to the type 1 error rate results, the power of all such construction methods and PS uses tended to align well with the theoretical estimates of power in the corresponding cluster randomized experiments. Notably, IPTW tended to have the highest power of all PS uses regardless of which method was used to construct the PS. However, practically the differences are negligible.

Table(3.65): Power of PS construction methods and PS uses

	Strata	Match	IPTW	Cov	Average
All	0.802	0.805	0.847	0.810	0.816
Impact	0.808	0.816	0.850	0.832	0.827
Stepwise	0.833	0.781	0.845	0.833	0.823

In summary, though the impact method attempts to minimize the variance of the estimator by removing those variables that have no effective relationship with the outcome, it does so conditional upon the potential bias reduction provided by those variables. As a result, regardless of the PS use, the simulations suggest that using the impact based construction does not comprise the power or type 1 error rates of the HLM estimator. Though an approach that solely minimizes bias or variance can detract from the quality of the estimator, using an approach that simultaneously minimizes the bias and the variance may improve the quality of the estimator without a corresponding inflation of type 1 error or decrease in power. For this reason, impact construction tends to capture the salient imbalances present among treatment groups in a manner that respects the ratio of the parameter estimate to its standard error. As a result, it potentially provides inferences similar to the all available variable PS construction approach. Such results emphasize the dual foci in estimating an effect: that of eliminating bias and that of minimizing variance. Sole focus on either can compromise the overall quality of the estimator. Table (3.66) displays the correspondence of inferences resulting from the HLM treatment effect estimator based on the three different PS construction methods. Table (3.66) shows that the all and impact methods make the same statistical inferences 97% of the time. Similarly, constructing the PS using the stepwise approach, one will tend to make statistical inferences similar to the all variables approach 95% of the time. Such results suggest that the impact method of constructing PS scores has less influence on statistical inferences than on point estimates. That is, though we see improved point estimates, the inferences resulting from impact construction of the PS tend to align with the all available variables approach.

Table(3.66): Probability of Obtaining the Same Inference by PS construction method

	All	Step	Impact
All	1		
Step	0.95	1	
Impact	0.97	0.95	1

Sensitivity of Thresholds to Observable Quantities

To provide insight as to how such thresholds may change as assumptions change, I conducted a variety of sensitivity analyses. These analyses were carried out by holding all other parameters constant at their default values and while varying a single parameter. As a result, we can examine how the thresholds are potentially related to the various parameters. I broke these analyses into two strands, sensitivity to directly observable quantities and sensitivity to those quantities that are not directly observable (though often testable).

In the first set of these sensitivity analyses, I focused on how such thresholds change as a function of relevant and observable quantities. In particular, I independently varied five parameters. First, I altered the probability of receiving the treatment, *p,* from the default of 0.50 to 0.90 as well as 0.10. These values were selected to understand how deviations from 0.5 might influence the thresholds. In particular, between the ranges of approximately 0.25 to 0.75, the logistic function is roughly linear. Beyond this range, the function takes on a more non-linear shape and potentially influences the thresholds in a different manner. Second, to encompass those studies that have little information as to the treatment assignment mechanism as well as those studies that are examining well studied treatments, I varied the Γ^* parameter or the correlation coefficient between the logit of the treatment assignment the observed predictors from 0.30 to 0.60 and 0.10. In a similar manner, I also varied the Δ^* parameter representing the correlation coefficient between the observed variables and the outcome. The Δ^* parameter was varied from its default of 0.70 to 0.30 and 0.80. Next, I altered the intra-class correlation, *ICC,* coefficient from

0.20 to 0.10 and 0.30. Such values align with those typically found in educational outcomes (Coe & Makoto, 2009). Further, such values additionally align with the subsequent application. Finally, I varied the group size, n_j, from 100 to 50 and 500. Fifty groups was selected as a lower bound as it has been suggested that 50 groups is generally a lower threshold from which to effectively use multilevel models (Maas & Hox, 2005; Moinedden, Matheson & Glazier, 2007). Further, the upper bound of 500 groups was selected since few educational studies exceed this many schools.

The results of the sensitivity of the thresholds to observed quantities are presented in Table (3.67). In each of the sensitivity analyses, the same essential pattern prevailed: the effect of the inclusion of a covariate in the PS model on the MSE of the HLM treatment effect estimator depended on the Γ- and Δ-relationships. Moreover, those covariates minimally related to the outcome but related to the treatment increased the variance of the estimator without decreasing bias. Further, the inclusion of covariates related only to the outcome decreased variance without affecting bias and the exclusion of those covariates related to both the outcome and the treatment yielded a biased and inefficient estimator. However, the alteration of the observable simulation quantities changed thresholds and, in some cases, changed the relative order of the thresholds.

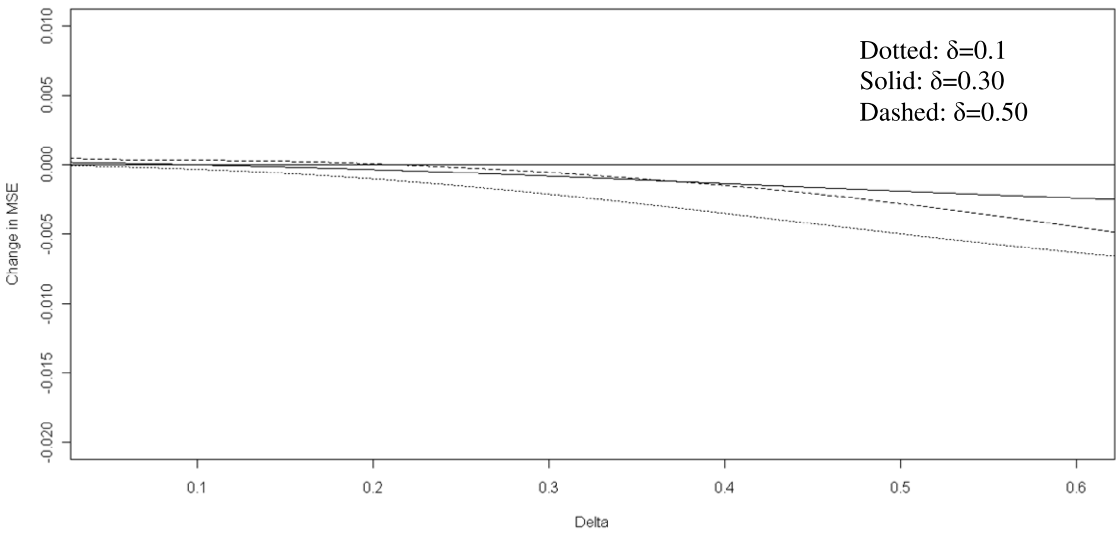
First, we saw that as we increase the probability of receiving the treatment, p, our thresholds tended to increase for the subclassification and covariance adjustments uses of the PS. However, in opposition the thresholds for the matching and IPTW uses of the PS tended to stay at magnitudes similar to that of a probability of 0.5. Next, as we increased the explanatory power of the observed covariates in the PS model for the treatment, Γ^*, we saw a consistent decrease in the magnitude of the threshold for all uses. That is, holding all other parameters constant including Δ*, as we increase the explanatory power of the covariates for the treatment, we should add covariates to the PS model at a lower Δ threshold. This aligns with the initial finding that in order to benefit from a variable's inclusion in the PS its Δ-relationship should exceed its Γ-relationship. With other words, if the covariates currently in PS model have a stronger Γ*-relationship than they do a Δ*-relationship, we need to recover the optimal balance and thus should include variables that would improve that balance in any way. Similarly, if we decrease Γ* while holding Δ* constant, the threshold becomes larger regardless of use. Paralleling such shifts, if we increase or decrease the relationship between the outcome and the covariates in the PS model (Δ*) while holding those covariates' relationship with the treatment (Γ*) constant, our thresholds will also increase or decrease respectively. That is, only a covariate with a relatively strong relationship with the outcome and smaller relationship with the treatment will likely improve our estimator once we have explained the majority of the variation in the outcome. Next, when altering the intra-class correlation, ICC, we are inherently shifting the proportion of each individual outcome the group is responsible for. Consequently, increasing the ICC tends to increase thresholds whereas decreasing the ICC tends to decrease the thresholds. In other words, the more the variation in the outcome the group is responsible for the earlier one should add group level covariates to the PS model. Finally, I considered the group size and its influence on the corresponding thresholds. Specifically, in this data the thresholds illustrated inverse relationships with the group size. That is, as one increases the group size, thresholds tend to decrease and vice versa. In other words, in this simulated data, larger group sizes allowed us to include more variables in the PS.

Table(3.67): Sensitivity of Thresholds to Observed Quantities

	Subclassification	Matching	IPTW	Cov
P=0.1	0.21	0.12	0.11	0.19
P=0.9	0.20	0.12	0.11	0.18
Γ=0.1*	0.20	0.14	0.12	0.18
Γ=0.6*	0.17	0.12	0.09	0.16
Δ=0.3*	0.17	0.10	0.09	0.17
Δ=0.8*	0.19	0.13	0.12	0.18
ICC=0.1	0.20	0.13	0.12	0.18
ICC=0.3	0.17	0.10	0.10	0.17
$n_j=50$	0.20	0.13	0.13	0.18
$n_j=500$	0.16	0.08	0.07	0.17

The results of the sensitivity of the thresholds to unobserved quantities are presented in Table (3.69). The first parameter I varied in this set of sensitivity analyses was the treatment effect, δ. In particular, because higher treatment effects tend to invoke increasing dependencies in the Γ- and Δ-parameters it is possible that higher treatment effects may change such thresholds (Pan & Frank, 2004). Take an artificial extreme example where we have an outcome being predicted by a treatment and confounding variable and the treatment is the variable of interest. If we allow the correlation between the outcome and treatment to be very high, e.g. 0.95, then if the confounder is highly correlated with the treatment (Γ-relationship) then the confounder will also be highly correlated with the (Δ-relationship). In particular, the Δ- and Γ-relationships are not independent and their product or impact tends to follow an approximate Beta distribution (Pan & Frank, 2004). As a result, such dependencies may skew thresholds. In the original experiment I used a true treatment effect of 0.3 whereas in the second I used 0.1 and 0.5. Such values in educational data typically align with small, moderate and large effect sizes (Cohen & Cohen, 1988). The results across uses of the PS were consistent in their direction but differed in magnitude. Specifically, as the treatment effect decreased or increased so did the corresponding threshold. Moreover, a decrease in treatment effect size tended to be associated with larger decreases in the threshold as compared to similar increases in the effect size with the exception of the covariance adjustment. In particular, excluding a variable when there is a small treatment effect tended to add a shrinking amount of bias to the estimator. That is, with a small treatment effect covariates tend to remove smaller amounts of bias. However when the treatment effect was large the inclusion/exclusion of a single covariate (depending on the Γ- and Δ-relationships) can add/subtract a considerable amount of bias. In turn such bias tends to add significantly to the MSE. Such relationships have extended effects as well. In particular, though higher effect sizes tend to have higher thresholds, the bias incurred by excluding a variable that exceeds the threshold (i.e. $\Delta >\psi$) grows much more rapidly with large effect sizes than with smaller effect sizes. For example, in Figure (3.68) if we were to exclude a variable in the PS model that has a Δ-relationship of 0.6, the increase in MSE will be larger if the true effect size is 0.5 (dashed) as opposed to 0.3 (solid). In a corresponding manner, identifying variables to omit from the PS when there is a small effect size is much more difficult as the difference in MSE between the small and large models becomes much smaller as illustrated by the gradual sloping and elongated curve (dotted) in Figure (3.68). In other words, when there are small effect sizes (e.g. 0.10) the amount of MSE that can be eliminated by using impact based construction decreases.

Figure(3.68): Example of Elongated Curves indicating earlier thresholds but less bias



Next, I considered a series of variables, interactions and higher order terms that may potentially be omitted in an analysis. First, I examined the sensitivity of the thresholds to an unmeasured group level variable with moderate relationships to both the outcome and the treatment. Subsequently, I assessed such sensitivities to excluded interactions and higher order terms among measured variables. In general, simulations indicated that thresholds would decrease when we have not taken into account an unmeasured variable. However, the change in thresholds when excluding an interaction is less clear but remains minimal. In addition, the scale of such decreased thresholds depends on the strength of the Γ- and Δ-relationships of the omitted variable.

Table(3.69): Sensitivity of Thresholds to Unobserved Quantities

	Subclassification	Matching	IPTW	Cov
δ=0.1	0.15	0.10	0.08	0.15
δ=0.5	0.20	0.13	0.13	0.19
Omitted W	0.18	0.10	0.08	0.16
Omitted Quadratic W	0.18	0.10	0.09	0.17
Omitted W_2W_3 interaction	0.19	0.13	0.13	0.18
Omitted ZW_3 interaction	0.16	0.13	0.11	0.18

Static vs. Dynamic Impact Construction

Turning to the changing nature of the thresholds, I present the results comparing the larger scale implementation of the static, dynamic, all and stepwise approaches are presented in Table (3.70). In particular, the table contrasts the approaches when 25 variables are available for the inclusion into the PS.

Table(3.70): MSE of PS Construction Methods and Uses when a Large Set of Potential Covariates Exists

	Subclassification	Matching	IPTW	Cov
All	0.0335	0.0561	0.0215	0.0321
Step	0.0539	0.0674	0.0216	0.0450
Dynamic Impact	0.0224	0.0342	0.0212	0.0223
Static Impact	0.0294	0.0523	0.0215	0.0300

We see that although prior simulations with only a few variables indicated that the static impact construction approach lowers MSE when we have appropriately identified the thresholds based on all parameters, not updating the thresholds as the parameters change throughout the PS model building process diminishes the benefit of using impact based construction. Alternatively, when using a dynamic impact based approach, we can retain much of the original benefit of using the impact based approach. Consequently, though identifying fixed thresholds for one's data is of some value, a superior approach might be to generate a series of thresholds corresponding to each step which allow you to evaluate the inclusion of a variable in the PS based on current estimates of the parameters.

Use of Proxies for Outcome

I conducted two further simulation experiments to assess the efficacy of using the pretest as a proxy for the outcome and using a cross validated approach to construct the PS. In particular, using simulated data similar to that above, I assessed the MSE of the treatment effect estimator when the PS model was built using the estimated thresholds and the Δ-relationships were estimated via the pretest or random subsample. The results of the experiment are presented in Tables (3.71) and (3.72).

Table(3.71): MSE of Treatment Effect Estimator when the PS is Constructed Using a Pretest Proxy Correlated at 0.70

	Strata			IPTW			Cov			Match		
	All	Step	Proxy Impact	All	Step	Proxy Impact	All	Step	Proxy Impact	All	Step	Proxy Impact
Bias	-0.0028	0.0074	-0.0029	-0.0036	-0.0128	-0.0016	-0.0021	0.0061	-0.0026	-0.0031	-0.0073	0.0211
SD	0.0374	0.0412	0.0361	0.0917	0.0938	0.0911	0.0412	0.0436	0.0374	0.0374	0.0374	0.0316
Var	0.0014	0.0017	0.0013	0.0084	0.0088	0.0083	0.0017	0.0019	0.0014	0.0014	0.0014	0.0010
MSE	0.0014	0.0018	0.0013	0.0084	0.0090	0.0083	0.0017	0.0019	0.0014	0.0014	0.0014	0.0014

Table(3.72): MSE of Treatment Effect Estimator when the PS is Constructed Using Cross Validation with 50% of the Sample

	Strata			IPTW			Cov			Match		
	All	Step	CV Impact	All	Step	CV Impact	All	Step	CV Impact	All	Step	CV Impact
Bias	0.0007	0.0028	0.0026	-0.0009	0.0035	0.0082	0.0019	0.0021	0.0021	-0.0013	0.0033	0.0021
SD	0.0548	0.0520	0.0500	0.1217	0.1241	0.1200	0.0566	0.0548	0.0510	0.0933	0.0721	0.0624
Var	0.0030	0.0027	0.0025	0.0148	0.0154	0.0144	0.0032	0.0030	0.0026	0.0087	0.0052	0.0039
MSE	0.0030	0.0027	0.0025	0.0148	0.0154	0.0144	0.0032	0.0030	0.0026	0.0087	0.0052	0.0039

Of immediate interest, I saw that for all uses and construction methods, building a PS on the basis of a proxy such as a pretest, provides a treatment effect estimator with lower MSE than doing so with cross validation. Using a cross validation based approach tends to nearly double MSE when compared to the MSE of a proxy based approach. However, in contrasting the cross validation and proxy based approaches, it is evident that the cross validation approach provides more bias reduction than the proxy approach. Constructing the PS by taking a random cross validation sample from the dataset should theoretically not bias estimates. Tables (3.71) and (3.72) provide empirical support for this hypothesis as cross validation had virtually no bias and considerably less than the proxy method. In contrast, constructing the PS based on a proxy measure that has less than perfect correlation with the outcome should theoretically bias estimates. However, such bias should be proportional to the strength of the proxy-outcome relationship. That is, as the strength of the proxy-outcome relationships increases the bias should decrease. Empirical support of this hypothesis is also evident in the results. Specifically, the results tended to illustrate residual bias that was frequently large relative to the cross validation results. Where the proxy based approach dominates the cross validation based approach is in terms of variance. Specifically, the proxy based approach tends to have about half the variance of the cross validation based approach. The proxy approach allows us to retain the full sample size whereas the cross validation approach reduces our sample size by half. As a result, theoretically, I hypothesized that the cross validation approach would have higher variance than the proxy approach. As a result, the two approaches outperform each other on different criteria. Accordingly the question then centers on the comparative contributions of the methods to the bias and variance of the estimator. In assessing the comparative contributions in the above tables, I noted that the magnitude of the bias is small relative to the variance. That is, the contribution of bias to the quality of the estimator is overshadowed by the contribution of variance.

Such results parallel the motivation behind this study in that in finite sample studies the variance of an estimator can play a role equal to or larger than the bias of an estimator. Such results also suggest careful attendance to such issues in applied settings. For example, researchers who have access to very large sample sizes and expect the study to be replicated, might decide to privilege the bias of an estimator. In such cases, it may be of better use to construct the PS using cross validation. Such an approach will likely offer a less biased or unbiased estimate of the true treatment effect. However, in moderate or small sized studies, researchers are perhaps more effective if they consider the bias variance tradeoff. Accordingly, if researchers have access to an outcome proxy such as the pretest, they should strongly consider constructing the PS using the proxy method. Use of such a method will likely improve the quality of estimation.

Further, the sub-optimality of such a step by step construction approach is also evident in Tables (3.71) and (3.72). In particular, because of the large set of covariates considered, only a step by step construction approach to the impact method was feasible. Though the MSE of the impact approach tends to offer estimators with less MSE than that of the all approach, we expected the MSE of the impact based approach to be considerably less than the MSE of the all approach. Such step by step procedures in conjunction with the proxy or cross validation tend to erode the reductions in MSE in practical model building. Despite such challenges, both the proxy and cross-validation approaches retained MSEs lower than stepwise in these simulations.

Discussion

The simulation experiments revealed that the model that best predicted treatment assignment did not yield a PS that minimized the treatment effect estimator in terms of MSE. Rather, the optimal model was the one that included only those variables whose contribution to the variance of the estimator was exceeded by its reduction in bias. This finding is consistent with the advice of Brookhart et al. (2006), Rubin and Thomas (1996) in that one should include in a PS model all variables thought to be related to the outcome, regardless of whether they are related to the treatment assignment. This study potentially advances such findings in that the impact based construction approach helps us quantify the magnitude of variable relationships to the outcome needed to effectively improve estimation. That is, such findings not only reinforce the necessity of including covariates unassociated to the treatment assignment but related to the outcome, but also quantify how strong that relation to the outcome needs to be. In the simulated data considered the strength of a covariate's relationship with the outcome needed to be at least half the size of the relationship with the treatment in order for its inclusion in the PS to decrease the MSE of the corresponding treatment effect estimator. Further, the impact based approach allows us to effectively and efficiently remove both systematic and non-systematic bias. In other words, for any given finite dataset, there are both actual confounders and empirical confounders. Whereas actual confounders lie along the causal pathway in the population, empirical confounders represent some small and usually statistically insignificant association or imbalance between covariates and the treatment assignment that arose by chance in a given sample. If such covariates additionally possess some nonzero relation to the outcome, then they are empirical confounders for that particular study. Though the systematic bias from such empirical confounders tends to zero as we take repeated samples and execute further studies, including such empirical confounders in a PS model for the study at hand removes the nonsystematic bias due to the chance association between the covariate and treatment assignment and improves the estimate for the study at hand. The removal of this nonsystematic bias tends to reduce the variance of an estimator thereby concentrating its density around its mean. This advice correlates with the theoretical finding that it can be advantageous to use an empirically estimated PS rather than the true population PS (Robins, Mark & Newey, 1992; Rosenbaum, 1987; Brookhart et al., 2006).

In efficiently removing such nonsystematic bias, the study indicated that if a covariate is related to the treatment assignment and has a relationship with the outcome less than the appropriate threshold, its inclusion in the PS model will increase the variance of the corresponding treatment effect estimator without decreasing bias. Such results come from the additional noise inserted into the estimated PS which causes an unnecessary inflation of the correlation between the estimated PS and the treatment assignment. In the case covariance adjustment or IPTW on the PS in an outcome model, including such a variable in the PS increases the covariance between the treatment and the PS increases and in turn increases the variance of the estimated treatment effect. Similarly, when context utilizing subclassification or matching on the PS in an outcome model, including covariates that do not meet the thresholds add noise to the estimated PS randomly misclassify or mismatch units with respect to important confounders.

In conclusion, the results presented in this study provide insight as how to identify variables to be included in the PS when it is used in conjunction with a HLM. Although such results are consistent with theoretical results the thresholds by which to include/exclude variables are dependent on the specification of the data generating process and the choice of different parameter values. Through sensitivity analysis, I varied the parameters that noticeably where the most relevant while fixing the probability distributions and other structural elements of the study. My findings and the analytical results presented by Rubin and Thomas (1997) and Robins et al. (1992) raise questions about the optimality of standard model building strategies for the construction of PS models, particularly in the setting of small to moderate sized studies. Model building algorithms that solely aim to create good predictive models of the treatment assignment neglect the duality of confounding and in turn often neglect the goals of the study. That is, as the ultimate goal of a study is often to effectively estimate a treatment effect, the purpose of utilizing the PS is to effectively and efficiently control for confounding so as to improve the properties of the treatment effect estimator rather than to predict the probability of receiving treatment. A variable selection criterion based exclusively on constructing the best possible treatment assignment model, be it best in prediction or in parsimony, will miss key variables related only to the outcome. Such an approach may miss important confounders that have a weak relation to the treatment but a strong relation to the outcome and detract from the quality of our estimate. Such auxiliary focus inherently limits the potential of observation data.

Example

As an example of how these methods might be applied to a substantive research question, I applied them to a study concerning school retention policies. That is, I asked what is the effect of allowing students to be retained, for any reason, on the average achievement level of a school? Holding a student back from advancing to the next grade is one potential approach to address persistently low performance by both students and schools (e.g. Roderick, Bryk, Jacobs, Easton & Allensworth, 1999). Advocates of school retention policies have suggested that when underperforming students are retained in a grade, classrooms tend to be more homogeneous (Shepard & Smith, 1988). As a result, teachers may be more able to manage their classrooms and target the learning needs of their students. Further, such management and targeted instruction within classrooms may potentially lead to increased average achievement across schools. However, in opposition some developmental psychologists have suggested that grade retention may constrain a student's cognitive and social development (Morrison, Griffith & Alberts, 1997). Further such retention may negatively affect a child's self-esteem and can increase their risk of dropping out in subsequent years (Roderick et al., 1999; Shepard, 1989).

Such school retention considerations have gained considerable attention in recent years, in part, as a result of initiative such as the No Child Left Behind act (NCLB) (NCLB, 2001). In particular, as schools have been under increasing pressure to be accountable for their student's performance and achievement they have assessed retention as a possible mechanism by which to increase progress (e.g Roderick et al., 2003; 2004). For instance, Chicago Public Schools and other schools throughout the country have adopted policies that end social promotion or in other words allow students to be retained if they have not satisfied the goals of the current grade level (Roderick et al. 2004). Although there have been a large number of studies and reviews concerning the effects of retention, the majority of such studies have focused on retention as an individual level treatment (Hong & Raudenbush, 2004; Jimerson, 2001; Holmes, 1989). Further, literature has tended to assume that the individual level treatment effect will generalize to the effect of school level retention polices. However, such aggregated assumptions may be misleading as the adoption of such policies by schools rather than individuals invokes a certain dependency in the data. That is, in assessing the effect of a school level policy such as retention we need to address the multilevel structure of the data. As a result, the effect of school retention policies on the average achievement of schools is inherently a multilevel question where students can be considered to be nested within schools.

To assess the effect of school level retention on school's average achievement, I used data from the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K). ECLS-K is a longitudinal dataset of a nationally representative sample of students, their families, teachers and the school they attend. Data on over 17,000 students were collected during three main time points: fall and spring of the kindergarten year and spring of the first grade year. Such a design has enabled researchers to consider the abilities and characteristics of schools and their students using multiple time points. Further, in this study schools reported whether it was the school's policy to allow retention of students based on any factor. That is, schools based on their collective attributes, including the distribution of students' abilities, selected their treatment status based on their perceived needs and anticipated benefits of assuming a social promotion or retention policy. Correspondingly, we can view the treatment, allowing retention, as being assigned at the cluster or school level. Within the context of ECLS-K, I considered the treatment to be attending a school that has allows retention during the second year of the study. The achievement metric of interest in this sample and this study was a mathematics score and a reading score which were both scaled using item response theory. Moreover, these scaled scores have been equated on the same scale to ensure comparability of the scores across time. Using these structures, I estimated the average effect of retention on math reading achievement separately using PS's constructed via the thresholds identified above in conjunction with two HLMs.

As the ECLS-K dataset represents observational data, the treatment assignment mechanism is unknown and needs to be inferred to approximate ignorability of the treatment assignment. The ECLS-K dataset contains a rich set of variables that, theoretically, may be predictive of which type of retention policy a school adopts. My approach to constructing the PS first involved identifying a suitable outcome proxy such as a pretest. In the ECLS-K dataset, the math and reading achievement from the spring of the first year have correlations exceeding 0.7 with the final math and reading achievement measures respectively. Accordingly, I used these measures as proxies for their respective outcomes in order to estimate each observed covariate's relationship with the outcome (Δ-relationship) and construct the PS in impact based manner. Further as such Δ- relationships take on a hierarchical structure, I utilized the weighted partial correlations. In particular, because such weights are based on the error variance (σ^2) and group variance (τ) which in turn depend on which covariates are considered at the student level, I first identified several student level covariates that have been historically related to student achievement. Although the inclusion of such covariates does not necessarily reduce the bias of the retention policy effect estimate, it likely improves the efficiency of the estimator. That is, because school retention policy is a school level treatment, ignorability of the treatment assignment can be achieved using school characteristics only. The first student characteristics I selected were the respective prior achievement measures from the first year. Second, I considered the gender of the student and third I considered two measures of socio-economic status.

To identify those observed variables that potentially influence a school's retention policy, I relied on prior literature. In particular, the covariates I considered for inclusion in the PS were based on those identified in Hong and Raudenbush (2006). The covariates included an extensive list of student and teacher aggregates and school variables including demographic characteristics, school characteristics and type, principal characteristics, school resources, neighborhood characteristics, assessment scores, home life and activities, parental involvement, physical and mental health, teacher characteristics, teacher and parent assessments, school learning experience and class composition (Appendix D). From this list I constructed the PS based on those variables whose Δ-relationship exceeded the threshold corresponding to its Γ-relationship suggested by the previous simulations. That is, I constructed the PS using only those variables which had a weighted partial correlation with the proxy outcome roughly twice as large as its weighted partial correlation with the logit of the treatment assignment.

Because the impact construction method attempts to construct meaningfully comparable groups for a given outcome, PS's may differ depending on the outcome of interest. That is, the impact based construction method decides which variables to include in the PS on the basis of the estimated Δ-relationships and Γ-relationships and the Δ-relationships change depending on the outcome. As a result, in assessing the effect of school policies that allow retention on the average math and reading achievement scores, I constructed two different PS's. The impact based PS for math identified 55 variables from which to predict the probability of adopting the school retention policy whereas the reading selected 54 (Appendix E). Though the PS models highly resembled each other, they did have a few differences. Specifically, whereas the PS for math achievement included the percent classified as gifted or talented in the school, the percent of students tutored in math, frequency of report cards and the regularity of unstructured play, the reading PS did not. In contrast, the reading PS included the severity of problems with gangs the school faces, a student's former teacher and his/her approaches to learning and the average number of gifted and talented within a classroom whereas the math PS did not.

To use these PS's to approximate the ignorability of the treatment assignment, I next employed stratification on the respective PS's. In particular, using the deciles of the logit of the PS I divided the schools into ten strata. After identifying the retention and non-retention schools within each stratum, I found no significant difference above the expected α-levels between the groups on the covariates using the general linear model. Tables (3.73) and (3.74) summarize the distribution of the logit of the PS for each stratum for the math and reading PS's.

Table(3.73): Balance of Logit of the Propensity Score for Retention Policy for Mathematics Achievement

	Retention Schools				Non-retention Schools	
Stratum	N	Mean	SD	N	Mean	SD
1	54	-0.33	0.18	65	-0.40	0.24
2	64	-0.03	0.05	54	-0.01	0.06
3	61	0.16	0.06	58	0.17	0.05
4	66	0.34	0.05	52	0.34	0.04
5	74	0.49	0.04	45	0.49	0.04
6	68	0.63	0.05	50	0.64	0.05
7	74	0.82	0.06	44	0.81	0.06
8	90	1.06	0.08	29	1.06	0.07
9	98	1.46	0.17	20	1.42	0.15
10	113	2.25	0.43	6	1.96	0.19

Table(3.74): Balance of Logit of the Propensity Score for Retention Policy for Reading Achievement

	Retention Schools				Non-retention Schools	
Stratum	N	Mean	SD	N	Mean	SD
1	67	-0.37	0.27	43	-0.32	0.17
2	63	0.01	0.06	46	0.00	0.06
3	73	0.19	0.05	36	0.20	0.05
4	68	0.34	0.04	41	0.34	0.04
5	73	0.47	0.04	36	0.48	0.04
6	88	0.64	0.05	21	0.62	0.05
7	81	0.83	0.07	28	0.85	0.05
8	89	1.10	0.07	20	1.08	0.07
9	99	1.48	0.15	10	1.42	0.12
10	102	2.23	0.41	8	2.49	0.56

Similar to the simulations above, I employed a hierarchical linear model with a random intercept to estimate the effect of a school retention policy. In particular, as prior literature and preliminary analyses have suggested that there was no systematic variation in the treatment effect across strata, I approximated the treatment using a hierarchical linear model (Hong & Raudenbush, 2006). The level one model concerning students was

$$Y_{ij} = \pi_{0j} + \sum_{p=1}^{P} \pi_p X_{pij} + \varepsilon_{ij} \tag{3.75}$$

where Y_{ij} represents math or reading achievement for the i^{th} student in school j, π_0 is the average student score adjusted for the student variables, X, and the corresponding coefficients, π_p, while ε_{ij} has a normal distribution with mean zero and variance σ^2. The school level model was

$$\pi_{0j} = \beta_{00} + \delta Z_j + \sum_{q=2}^{10} \beta_{0q} S_{qj} + r_{0j} \tag{3.76}$$

where β_{00} is the average adjusted achievement for school, δ is the average effect of school policies which allow retention, β_{0q} is average effect of strata S_{qj}, on adjusted achievement and r_{0j} is the random effect of school j and has a normal distribution with mean zero and variance τ_π. Specifically, I used strata indicators and the school level I used the student level covariates mentioned above to address any remaining intra-stratum bias and increase the efficiency of the estimator.

The results of the analyses are presented in Tables (3.77) and (3.78). Using an impact based PS in conjunction with a hierarchical linear model suggested that on average a school's retention policy was not significantly associated with an increase or decrease in the school's overall achievement.

Table(3.77): Average School Retention Policy Effect on Math Achievement

Fixed Effect	Estimate	SE	t
Average math achievement in non-retention school (β_{00})	23.27	0.34	69.05

School retention policy effect (δ)	-0.01	0.18	-0.04
Math achievement from spring of first year (π_1)	0.61	0.01	73.75
Math achievement from fall of first year ((π_2)	0.16	0.01	15.89
Female (π_3)	-0.30	0.09	-3.27
SES (π_4)	0.60	0.09	6.69
Mother's education level (π_5)	0.07	0.03	1.92
PS stratum 2 (β_{02})	0.40	0.34	1.19
PS stratum 3 (β_{03})	0.43	0.33	1.29
PS stratum 4 (β_{04})	0.31	0.33	0.93
PS stratum 5 (β_{05})	0.27	0.34	0.82
PS stratum 6 (β_{06})	0.12	0.34	0.36
PS stratum 7 (β_{07})	0.33	0.35	0.95
PS stratum 8 (β_{08})	0.30	0.36	0.83
PS stratum 9 (β_{09})	-0.74	0.36	-2.03
PS stratum 10 (β_{010})	-0.68	0.38	-1.82

Table(3.78): Average School Retention Policy Effect on Reading Achievement

Fixed Effect	Estimate	SE	t
Average reading achievement in non-retention school (β_{00})	26.91	0.70	38.28
School retention policy effect (δ)	-0.01	0.11	-0.14
Reading achievement from spring of first year (π_1)	0.70	0.02	45.70
Reading achievement from fall of first year ((π_2)	0.35	0.02	18.68
Female (π_3)	2.60	0.17	15.34
SES (π_4)	1.86	0.17	11.19
Mother's education level (π_5)	0.24	0.06	3.71
PS stratum 2 (β_{02})	0.40	0.74	0.55
PS stratum 3 (β_{03})	0.48	0.75	0.64
PS stratum 4 (β_{04})	-0.11	0.74	-0.15
PS stratum 5 (β_{05})	0.35	0.74	0.47
PS stratum 6 (β_{06})	0.34	0.76	0.44
PS stratum 7 (β_{07})	-0.55	0.75	-0.72
PS stratum 8 (β_{08})	-1.14	0.77	-1.49
PS stratum 9 (β_{09})	-1.59	0.77	-2.07
PS stratum 10 (β_{010})	-4.66	0.79	-5.92

More specifically, my estimates suggested that for both math and reading there was virtually no statistical or practical difference between those schools that allow student retention and those schools that did not allow it. Further, earlier simulations exploring the impact based PS construction method suggest that such estimates tend to be more concentrated around the true effect of such retention policies when compared to estimates based on using all the available variables or stepwise construction of the PS. However, such focus on school policy without a corresponding focus on student level retention risk factors and the differential likelihoods and criteria by which schools retained children, may not present a holistic picture of retention (Hong & Raudenbush, 2006).

CHAPTER IV

Robustness of Causal Inferences in Binomial Regression Models

Introduction

In education observational studies that assess the effect of a treatment, researchers frequently rely on measuring and adjusting for potential confounding variables to provide sound estimates of the treatment effect. However, in such studies it is virtually impossible to measure all potential confounding variables. Further, researchers are frequently unable to measure all variables that are hypothesized to be confounded with the treatment for a given outcome. Yet, without proper adjustment for all confounding variables, common estimators may present biased estimates of the treatment. One alternative is to employ randomization of the treatment. In a randomized experiment the treatment assignment mechanism is known to be random and unrelated to subject's potential outcomes as well as pretreatment characteristics. Randomization is an impartial method to construct group membership in that it ensures that any and all pretreatment differences are only by chance. Group construction in such a manner implies that chance imbalances between groups tend to zero as the number of subjects approaches infinity. This property provides us with a simple and known error distribution that converges. Consequently, in comparing the outcomes of the different groups after treatment, we can quantify our confidence about estimates of the treatment effect. However, such benefits are often mitigated by limitations of randomized experiments. For instance, there are numerous situations where one can not feasibly randomize treatments for ethical or financial restrictions. Further, even when randomization may be feasible, the practical implementation can be contaminated as seen in studies such as the Tennessee class size experiment (Krueger, 1999).

Another alternative when one can not measure an exhaustive list of confounding variables is to understand the sensitivity of one's inferences to an unmeasured confounding variable. One such general framework for assessing the robustness of an inference is a sensitivity analysis (e.g. Rosenbaum, 1995; Frank, 2000). Through this general structure researchers have, in various ways, attempted to quantitatively evaluate the sensitivity of their claims to an omitted variable by assessing how much imbalance on an unknown covariate it would take to invalidate their inference. In this study, I extended Frank's (2000) Impact Threshold of a Confounding Variable (ITCV) on a regression coefficient concept to encompass binomial regression models (BRMs).

Though inferences from well implemented randomized studies have high internal validity they often offer lower external validity than observational studies. For example, though the Tennessee class size experiment indicated that class size reduction was associated with positive achievement gains in Tennessee, similar implementation in California did not result in such gains (e.g. California Legislative Analyst's Office, 1997). In other words, the sample and practical constraints of implementing a randomized experiment can often impede the goals of the study.

In a similar manner observational studies offer several advantages and drawbacks. For example, large scale observational studies tend to be significantly less expensive to implement. As a result, representative samples from the populations one hopes to draw inferences on can be selected. In contrast to such high external validity, observational studies may have less internal validity. In particular, researchers do not have the benefit of randomly assigning subjects to groups. In this type of study, where there is an absence of random assignment, groups are potentially imbalanced on pretreatment characteristics that may or may not influence the outcome. As a result, differences in outcomes may be a function of pretreatment imbalances between the groups rather than a treatment effect. However, if such imbalances are accurately observed and measured, imbalances can be appropriately adjusted for by analytic methods such model based covariance adjustments. Although such methods offer accurate estimation of treatment effects, their accuracy is tempered by the threat of an unmeasured baseline imbalance between the groups. In such a case, and potentially in every observational study, two subjects who are identical on measured characteristics may have an unequal probability of choosing or being assigned to the treatment or control groups due to differences on an unmeasured characteristic.

Recognizing the potential for imbalance in observational or quasi-experimental studies, I shift attention to the pragmatic question of how much imbalance there must be to invalidate an inference. That is, it is crucial to understand the quality of one’s inferences including the general sensitivity of one’s inferences to an unmeasured variable in such studies. Frank (2000) defines the ITCV in the linear regression model for continuous outcomes estimated by ordinary least squares (OLS). However, in many disciplines including education, there are discrete outcomes which have nonlinear relationships with treatments. Specifically, in this context I focus inquiry on the BRM as represented through generalized linear models (GLMs) (McCullagh & Nelder, 1983). Frank (2000) considers the ordinary least squares (OLS) regression model

$$y = \beta_0 + \beta_1 Z + \varepsilon \tag{4.1}$$

where *Z* is a treatment or variable of interest, with a continuous or discrete distribution, of which we are interested in estimating its coefficient, β_1. Using the (OLS) framework estimator (4.1) provides an unbiased, efficient estimate of β_1 given that the conditional expectation (on *Z*) of ε is zero. Implied by this assumption is there are no omitted variables correlated with both *Z* and ε or in other words no confounding variables. As confounding can be characterized through correlation in linear models, Frank then asks, given β_1 is statistically significant in model (4.1), for a given outcome-treatment correlation (hereafter referred to as the Φ-relationship), how large must the product of an unmeasured confounding variable's, *U*, correlations with the outcome (hereafter referred to as the Δ-relationship) and the predictor of interest (hereafter referred to as the Γ-relationship) be to make β_1 insignificant in

$$y = \beta_0 + \beta_1 Z + \beta_2 U + \varepsilon \tag{4.2}$$

This approach quantifies how much imbalance on an unobserved confounding variable is needed to invalidate an inference concerning β_1. Similar to other sensitivity analyses, the measure can quantify under what circumstances β_1 is insensitive to unobserved variables. Accordingly, when treatment assignment is well balanced and unrelated to other factors influencing the outcome and the test statistic is large, to invalidate the inference resulting from (4.1), the correlations of the unobserved variable with the outcome and with *Z* would have to be large. To quantify this sensitivity in terms of correlations Frank (2000) writes β_1, its standard error and *t*-ratio in terms of correlations as

$$\widehat{\beta}_1 = \left(\frac{s_y}{s_z}\right)\left(\frac{r_{yz} - r_{yu} r_{zu}}{1 - r_{zu}^2}\right) \tag{4.3}$$

$$se(\hat{\beta}_1) = (\frac{s_y}{s_z}) * \sqrt{\frac{1 - (\frac{r_{yz}^2 + r_{yu}^2 - 2r_{yz}r_{yu}r_{zu}}{1 - r_{zu}^2})}{n - q - 1} * \frac{1}{1 - r_{zu}^2}} \tag{4.4}$$

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{(r_{yz} - r_{yu}r_{uz})}{\sqrt{\frac{1 - r_{uz}^2 - (r_{yz}^2 + r_{yu}^2 - 2r_{yz}r_{yu}r_{zu})}{n - q - 1}}} \tag{4.5}$$

Here $s_.$ is the standard deviation, n is the sample size, q is the number of predictors excluding the intercept, r_{yu} represents the Δ-relationship and r_{uz} represents the Γ-relationship and in general $r_{..}$ is the appropriate zero correlation based on

$$r_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_x s_y} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_x s_y} = \frac{n\sum X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} + \sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}} \tag{4.6}$$

Using these expressions Frank (2000) then derives the threshold for invalidating an inference as a function of the product of the Δ- and Γ-relationships based on the *t*-ratio inference statistic. In particular, using the product of the Δ-and Γ-relationships he derives the minimal confounder correlations necessary to invalidate the inference on β_1 in (4.1). As a result, using only the known data, researchers can assess the sensitivity of their inference in (4.1) to an unmeasured confounding variable in terms of a common social science metric, correlation.

Research Questions

In this study, I investigated the concepts and methods of estimating the impact threshold of an unmeasured confounding variable in BRMs estimated by maximum likelihood (ML). Using the binomial model with a logit link, I examined two challenges to extending the impact threshold to binomial regression models. The first is defining a framework within BRMs such that correlation represents a useful and meaningful statistic. The second is developing a method to approximate such thresholds. To this end I focused my research on the following questions:

1. How can we extend the framework of Impact Threshold of a Confounding Variable to BRMs?
2. Given the iterative nature of estimation and non-deterministic nature of point estimates by summary statistics in BRMs, how shall we define the Impact Threshold of a Confounding Variable in binomial regression models? What are the key extensions needed to replace ITCV?
3. Can the distribution of inference statistics resulting from the impact of a confounding variable (*test*-statistic) be approximated well? If so, what is the approximate distribution? Does such a distribution provide an informative and practically useful (narrow) range for the impact of the confounding variable on the *test*-statistic of the treatment?
4. Is the average *test*-statistic a monotonic function of the impact?
5. How does the ITCV change for multiple predictors?

Theoretical Framework

In models that consider continuous outcomes and homogeneity of variance, the standard linear model using the OLS estimator is the most common method for estimating parameters of interest. In OLS, the vector of parameters can be estimated in closed form by the least squares estimator which is equivalent to

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{\mathbf{-1}}(\mathbf{X}^{\mathbf{T}}\mathbf{Y}) \qquad (4.7)$$

where $\boldsymbol{\beta}$ represents the *p x 1* vector of parameters of interest, **X** is the *n x p* design matrix which includes a vector of ones for the intercept and **Y** is the *n x 1* matrix of responses. In such circumstances it can be shown that the OLS estimator is also the ML estimator when errors are normally distributed. In contrast, models that consider binary outcomes often exhibit heterogeneity of variance as the variance is linked to the mean. In such situations, the OLS estimator is not a desirable estimator as it is inefficient and its assumptions such as homoskedasticity are directly violated. In such circumstances the BRM is the most common model. In general, BRMs are part of a class of models known as generalized linear models (GLMs) which can be defined by specifying two components: the distribution of the response and the link function. In the first component, the distribution of the response is necessarily from the *j*-parameter exponential family of distributions of general form

$$f(y;\boldsymbol{\theta}) = \exp\{\Sigma_{\mathrm{j}}\, A_j(\boldsymbol{\theta})\, B_j(y) + C(y) + D(\boldsymbol{\theta})\} \tag{4.8}$$

where $\boldsymbol{\theta}$ is the estimand, $A_j(\boldsymbol{\theta})$ and $D(\boldsymbol{\theta})$ are functions of $\boldsymbol{\theta}$ alone and $B_j(y)$ and $C(y)$ are well-behaved functions of the data alone. The second component, the link function, specifies the relationship between the outcome and the parameters and is subsequently discussed. Although, the relationship between the outcome and the parameters is nonlinear with discrete outcomes, GLMs consider the transformed outcome to be linear in the parameters.

Conceptually, BRMs represent the outcome through some monotone continuous and differentiable function g(y) and model the expected response as a function of the covariates rather than the actual response. The transformation of the outcome is achieved through a linearized version of the link function using a first order Taylor expansion. As a result BRMs utilize a link function, η, such that η = g(μ) and μ = E(Y). Using these relations BRMs linearize g(y) using a one step Taylor expansion as follows:

$$\begin{aligned} g(\mathrm{y}) &\approx g(\mu)+(\mathrm{y}-\mu)g'(\mu) \\ &=\eta+(\mathrm{y}-\mu)\frac{\partial\eta}{\partial\mu} \\ &\equiv Y^{(i)} \end{aligned} \tag{4.9}$$

In estimating the parameters of interest in BRMs, the ML estimator is the most common estimator[1]. Maximum likelihood theory conceptualizes the likelihood as a function of the parameters given the observed data and seeks to find the value of the parameter(s) that gives the largest probability to the observed data or in other words maximize the likelihood function. Resulting estimators have well studied properties and are asymptotically unbiased, efficient and normally distributed. Although ML estimators have desirable theoretical properties they frequently can not be estimated in closed form and require numerical methods to be maximized. For example, let $Y_1, Y_2,..., Y_n$ be independent and identically distributed binomial observations with probability π_i, that is $\boldsymbol{Y}$~ binomial (n, π). Moreover, assume we build a model using, say, two covariates and an intercept where Z is the treatment status for each covariate class, and U is the value of a confounding variable for each covariate class. Using the canonical logit link and three parameters β_0, β_1, and β_2 we have the model

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 Z + \beta_2 U \tag{4.10}$$

with corresponding log likelihood

$$l = \Sigma_\mathrm{i} \ln\binom{n}{Y_i} + \beta_0\Sigma_\mathrm{i}\, \mathrm{Y_i} + \beta_1\Sigma_\mathrm{i}\, \mathrm{Z_iY_i} + \beta_2\Sigma_\mathrm{i}\, \mathrm{U_iY_i} + \Sigma_\mathrm{i} \ln(1+\exp[\beta_0+\beta_1\mathrm{Z_i}+\beta_2\mathrm{U_i}]) \tag{4.11}$$

The corresponding score equations are

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} \left(Y_i - \frac{\exp[\beta_0 + \beta_1 Z_i + \beta_2 U_i]}{1 + \exp[\beta_0 + \beta_1 Z_i + \beta_2 U_i]} \right) = 0 \qquad (4.12)$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} \left(Z_i Y_i - \frac{Z_i \exp[\beta_0 + \beta_1 Z_i + \beta_2 U_i]}{1 + \exp[\beta_0 + \beta_1 Z_i + \beta_2 U_i]} \right) = 0$$

(4.13)

$$\frac{\partial \ell}{\partial \beta_2} = \sum_{i=1}^{n} \left(U_i Y_i - \frac{U_i \exp[\beta_0 + \beta_1 Z_i + \beta_2 U_i]}{1 + \exp[\beta_0 + \beta_1 Z_i + \beta_2 U_i]} \right) = 0$$

(4.14)

In such a model, closed form estimates for $\boldsymbol{\beta}$ are not available and maximizing the likelihood requires numerical optimization. As a result, sensitivity analyses can not generally be estimated using closed form. In BRMs the most common numerical method to maximize the likelihood is Fisher scoring which is equivalent to iteratively re-weighted least squares (IRWLS) (McCullagh & Nelder, 1983).

Using this equivalence, we can recast ML estimation in BRMs as an iterative least squares regression model. Utilizing the above transformation, (4.9), we can summarize IRWLS algorithm by the following steps (where *i* indicates the *i*th iteration):

(1) Set estimates of $\hat{\eta}_i$ and $\hat{\mu}_i$

(2) Form the adjusted linearized dependent variable $Y^{(i)} = \eta_i + (\text{y} - \mu_i)\left(\frac{\partial \eta}{\partial \mu} \mid \eta_i\right)$

(3) Form the weights $w_i^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 \mid \hat{\eta}_i V(\hat{\mu}_i)$

(4) Estimate $\boldsymbol{\beta}^{(i)}$ using the weighted least squares estimator

(5) Using $\boldsymbol{\beta}^{(i)}$ calculate the new values of the linear predictor, $\eta^{(i+1)}$ and $\mu^{(i+1)}$

(6) Repeat steps 2-5 until a convergence criterion is reached

Here the weights are based on the variance of $Y^{(i)}$ in (4.9) which is

$$\widehat{\text{var}}(Y^{(i)}) = \left(\frac{\partial\eta}{\partial\mu}\right)^2 V(\hat{\mu}) = \frac{1}{w} \qquad (4.15)$$

where $V(\cdot)$ is the variance function, μ_j is the mean or predicted response for observation j, and w are the weights for each observation which reflect the differing levels of uncertainty. For BRMs we have

$$V(\hat{\mu}) = \frac{\hat{\mu}(1-\hat{\mu})}{n} \qquad (4.16)$$

To simplify notation we write the scoring algorithm for the i^{th} iteration in matrix form as

$$\boldsymbol{\beta}^{(i)} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}^{(i)}\mathbf{X})^{-1}(\mathbf{X}^{\mathbf{T}}\,\mathbf{W}^{(i)}\mathbf{Y}^{(i)}) \qquad (4.17)$$

where $\boldsymbol{\beta}$ is vector of parameters, **X** is the design matrix, **Y** is the *adjusted* outcome vector and **W** is the weight matrix such that

$$\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{nn} \end{pmatrix} \quad (4.18) \qquad \text{and} \qquad \mathbf{Y} = \begin{pmatrix} \mathrm{y}_1 \\ \vdots \\ \mathrm{y}_n \end{pmatrix} \quad (4.19)$$

Moreover, $\mathbf{Y}^{(i)}$ and the diagonal elements of **W** (e.g. w_{jj}) represent the adjusted dependent variable and the inverse of the variance of the adjusted dependent variable on the i^{th} scoring iteration in accordance to steps (2) and (3) above, respectively.

Methods

Extending the ITCV to BRMs

BRMs do not specifically consider outcomes to have linear relationships with covariates but rather define the outcome to be linearly related to the covariates through the link function, η. That is, the link function is now a linear function of the covariates rather than the actual outcome itself. As a result, the utility of correlations in defining the

ITCV in BRMs now rests in using the link function (or linearized version of the adjusted dependent variable). More specifically, defining the ITCV in BRMs in terms of correlation requires us to consider the link function as the dependent variable rather than the original, untransformed outcome. For instance, in estimating correlations with the outcome for the binomial model with the logit link, we now consider the correlation between the predictor of interest and transformed random variable *y*, the logit of the observed binomial proportions. That is, we consider the correlation with the logit function of the binomial observed proportions, $\hat{p}$, such that $\eta = \ln(\frac{\hat{p}}{1-\hat{p}})$. In other words, we use the sample correlation, $r_{x\eta}$, between the outcome (logit ($\hat{p}$)) and predictor of interest is

$$r_{\mathrm{x}\eta} = \frac{\sum(X_i - \bar{X}\ )(\eta_i - \bar{\eta})}{(n-1)s_{\mathrm{x}}s_{\eta}} = \frac{\sum X_i\eta_i - n\bar{X}\bar{\eta}}{(n-1)s_{\mathrm{x}}s_{\eta}} = \frac{n\sum X_i\eta_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n}\eta_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} + \sqrt{n\sum_{i=1}^{n}\eta_i^2 - (\sum_{i=1}^{n}\eta_i)^2}} \qquad (4.20)$$

where η is the natural logarithm of the estimated odds as oppose to the standard outcome, *y*. Further, in the case where the logit is undefined as a result of no observed successes, the logit of the outcome is adjusted by adding a half of success to the number of successes and totals (Agresti, 1996). This conceptual adjustment of the outcome from proportion to logit is a first step in framing the ITCV in BRMs.

The equivalence of ML estimation and IRWLS in BRMs provides a natural pathway to extend the concept of the ITCV to BRM. However, basing such estimation on zero order correlations as we did in the OLS case, takes on a more complex form resulting from the iterative nature of estimation, the distribution of the outcome and the unequal variance of the responses. In particular, there are several notable changes including the outcome variable, the unequal variance of observations, the constraints of the dispersion parameter, the exchange of inference theory, and the numerical estimation of the parameters.

First, I attend to the weighted nature of IRWLS and, for the moment, assume the likelihood is maximized using a single iteration thereby making it equivalent to weighted least squares (WLS). Whereas OLS implicitly assumes the weight matrix, **W**, is the identity matrix as observations have equal variance, BRMs must explicitly consider and identify the potentially non-equal weights between observations. To address this variation, we now necessarily need to estimate *weighted* zero order correlations to ensure our estimator is efficient. More specifically, in OLS we estimated the correlation using the unbiased estimator:

$$r_{\mathrm{xy}} = \frac{\sum (X_i - \bar{X}\ )(Y_i - \bar{Y})}{(n-1)s_{\mathrm{x}}s_{\mathrm{y}}} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_{\mathrm{x}}s_{\mathrm{y}}} = \frac{n\sum X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} + \sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}} \qquad (4.21)$$

Whereas in WLS (or IRWLS) we need estimate the correlations by the unbiased estimator:

$$r_{w_{\mathrm{xy}}} = \frac{\sum w_i (X_i - \bar{X}_w)(Y_i - \bar{Y}_w)}{(1 - \sum w_i^2) s_{w_{\mathrm{x}}} s_{w_{\mathrm{y}}}} \qquad (4.22)$$

where $\bar{X}_w = \sum w_i X_i$, $\bar{Y}_w = \sum w_i Y_i$, (4.23)

$$s^2_{\mathrm{w\,x}} = \frac{\sum w_i}{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum w_i^2} - \sum w_i (\mathrm{X}_i - \bar{X}_w)^2 \qquad (4.24)$$

and

$$s^2_{\mathrm{w\,y}} = \frac{\sum w_i}{(\sum_{i=1}^{n} w_i)^2 - \sum w_i^2} - \sum w_i (Y_i - \bar{Y}_w)^2 \qquad (4.25)$$

and without loss of generality we assume $\sum w_i = 1$. As a result, using the sample weighted correlation estimator, $r_{w_{..}}$, rather than the usual sample correlation, $r_{..}$, allows us to estimate weighted least squares regression coefficients using (4.3). That is, the WLS estimator of β_1 in (4.10) is equivalent to the following function of weighted correlations

$$\beta_1^{wls} = [(\mathbf{X}^\mathbf{T}\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^\mathbf{T}\ \mathbf{W}\mathbf{Y})]_{(2,1)} = \left(\frac{s_{w\eta}}{s_{\mathrm{w\,z}}}\right)\left(\frac{r_{w\eta\mathrm{z}} - r_{w\eta\mathrm{u}}\, r_{\mathrm{wzu}}}{1 - r^2_{\mathrm{wzu}}}\right) = \beta_1^{wls} \qquad (4.26)$$

Here, I have redefined the correlation to reflect the inherent unequal variance in the observations. In the same manner I adjust the standard error of β_1 in that we must now replace each correlation, $r_{..}$, with the corresponding weighted correlation, $r_{w_{..}}$. In addition, standard errors in WLS for continuous outcomes allow the dispersion parameter to take any positive value less than infinity. However, in BRMs the dispersion parameter is constrained to be one by assumption. As a result, the standard errors need to be scaled by the estimate of the dispersion parameter, σ. That is, in OLS models (e.g. (4.2)) the

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = (\mathrm{X}^\mathrm{T}\mathrm{WX})^{-1}(\hat{\sigma}^2) \qquad (4.27)$$

whereas in BRMs the

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) = (\mathrm{X}^\mathrm{T}\mathrm{WX})^{-1} \qquad (4.28)$$

Therefore, the OLS standard error of β_1 from (4.4) misestimates the standard error of β_1 in (4.10) by a factor of σ. Consequently, we need to rescale the standard error of the correlation estimator by σ. Thus we modify (4.4) such that

$$se(\widehat{\beta}_1) = \frac{(\frac{s_{w_y}}{s_{w_z}}) * \sqrt{\frac{1-(\frac{r^2_{w_{yz}} + r^2_{w_{yu}} - 2r_{w_{yz}} r_{w_{yu}} r_{w_{zu}}}{1-r^2_{w_{zu}}})}{\text{n - q - 1}} * \frac{1}{1-r^2_{w_{zu}}}}}{\sigma} \tag{4.29}$$

Finally, as we are estimating the parameters with ML, the parameters no longer have an exact *t*-distribution. Maximum likelihood theory indicates that ML estimates are asymptotically normal implying that the distribution of our parameter estimates is best approximated by the *z*-distribution rather than the *t*-distribution. Accordingly, inferences about the parameter of interest are now based on

$$z(\widehat{\beta}_1) = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)} = \frac{(r_{w_{yz}} - r_{w_{yu}} r_{w_{zu}})\sigma}{\sqrt{\frac{1-r^2_{w_{zu}} - (r^2_{w_{yz}} + r^2_{w_{yu}} - 2r_{w_{yz}} r_{w_{yu}} r_{w_{zu}})}{n-q-1}}} \tag{4.30}$$

Defining the ITCV for BRMs

Parameter estimates based on the principles above using only a single iteration would amount to WLS estimates rather than ML estimates. In order to obtain ML estimates we would need to utilize an iterative process that recycles the WLS parameter estimators but updates the adjusted dependent variable and observation weights until the parameter estimates converge. Tracking this part of the extension proves difficult as summary statistics such as correlation are not sufficient as they are not necessarily one-to-one functions of the ML estimates (Rao, 1952; Berkson, 1957). To see this, consider

ML estimation of β_1 in (4.10). By a property of the exponential family of distributions, if we write the probability density function as in (4.8), then

$$\boldsymbol{T} = \Big(\sum_{j=1}^{n} B_1(\mathrm{y}_i), \sum_{j=1}^{n} B_2(\mathrm{y}_i), \ldots, \sum_{j=1}^{n} B_k(\mathrm{y}_i)\Big) = (t_1, t_2, \ldots, t_k) \tag{4.31}$$

T is a (minimal) sufficient statistic and is thus (jointly) sufficient for $\boldsymbol{\beta}$ (Garthwaite, Jolliffe & Jones, 2002, p. 30). As a result, the corresponding minimal sufficient statistics for $\boldsymbol{\beta}$ in (4.10) are:

$$\text{for } \beta_0\text{:} \quad \sum_i Y_i \tag{4.32}$$

$$\text{for } \beta_1\text{:} \quad \sum_i Z_i Y_i \tag{4.33}$$

$$\text{for } \beta_2\text{:} \quad \sum_i U_i Y_i \tag{4.34}$$

Now, note that the relevant correlations can be written as

$$r_{\mathrm{yz}} = \frac{\sum Y_i Z_i - n\overline{Y}\,\overline{Z}}{(n-1)s_{\mathrm{y}}s_{\mathrm{z}}} \tag{4.35}$$

and

$$r_{\mathrm{yu}} = \frac{\sum Y_i U_i - n\overline{Y}\,\overline{U}}{(n-1)s_{\mathrm{y}}s_{\mathrm{u}}} \tag{4.36}$$

where $s_{.}$ is the standard deviation. Without loss of generality, assume that the data sets have been standardized such that each variable has a mean of 0 and standard deviation of 1. As a result, we see that the correlations reduce to their respective minimal sufficient statistic scaled by a factor of *1/(n-1).* Minimal sufficient statistics are not unique in that any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic (e.g. Casella & Berger, 2002, p. 282). In addition to $\boldsymbol{T}$ being minimal sufficient, note that (4.35) and (4.36) are also minimal sufficient when the variables are standardized.

Although quantities such as correlation are theoretically sufficient for the ML estimates of the parameters in (4.10), practically they are not sufficient for estimation. As parameter estimates from ML estimates only reduce to closed form solutions in trivial models, numerical optimization, such as Fisher scoring above, must be employed. In general, such methods require more than a summary statistic such as correlation to identify parameter estimates as the weights of observations tend to be unknown. Consequently, correlation and other statistics do not uniquely identify the parameter estimates or inference statistics. As a result, different confounding variables with identical Δ- and Γ-values will alter the parameter estimates and test statistics in different ways. This concept is evident in many canonical sensitivity analyses as they result in intervals of inference statistics rather than a single inference statistic (e.g. Rosenbaum, 1995).

In order to address the second research question, I redefine the ITCV in BRMs by replacing the impact threshold with the average impact threshold of a confounding variable (AITCV). More specifically, I define the AITCV as the Δ-Γ product which reduces the average *test*-statistic of the treatment to some critical level (e.g. $z = 1.96$). That is, because specifying the Γ- and Δ-relationships does not uniquely identify a new *test*- statistic for the treatment but rather a range of test-statistics, I define the AITCV as the Γ- and Δ-relationships that produce an average *test*-statistic equal to some critical value (e.g. 1.96). For instance, to identify the average impact of a confounding variable

(AICV), one would first consider all confounding variables with a given Γ- and Δ-value and estimate how controlling for each confounding variable altered the original test statistic in (4.10). Subsequently, one would take the average of those *test*-statistics and call this the AICV. Similarly, to identify the AITCV, one would repeat this process until the average test statistic matched the critical value (e.g. 1.96 for α-level of 0.05 in a *z*-distribution).

Approximating the Distribution of the Inference Statistics

A task in understanding the ability of such a threshold to summarize the impact of a confounding variable is to characterize the distribution of plausible *test*-statistics given the speculation parameters (i.e. Γ and Δ). In particular, as the ITCV framework defines the threshold with respect to inference, it is central to approximate the distribution of the resulting inferential statistics (i.e. *test*-statistic). That is, each confounding variable with fixed Γ- and Δ-relationships (e.g. correlations) may change the original *test*-statistic by differing amounts depending on the exact data points. As a result, specifying only the Γ- and Δ-relationships and controlling for the confounding variable does not determine a single new *test*-statistic but rather gives us a range of *test*-statistics. A crucial criterion from which to judge the ability of the AITCV approach to summarize this range is the dispersion and density of the *test*-statistics corresponding to the speculation parameters (i.e. Γ and Δ). Further, in approximating such a distribution, a central task is how to identify and appropriately consider the hypothesized relationships of potential confounding variables in the parameter space. Through Monte Carlo simulation analysis I randomly sampled from the parameter space of potential confounding variables and re-estimated the treatment effect using (4.10). Repeating this process, I used the new resulting sample *test*-statistics of the treatment to approximate their distribution and thus consider the impact of confounding variables with various Δ- and Γ-relationships. To evaluate the ability of the AITCV approach in summarizing the impact of a confounding variable in BRMs, I assessed this distribution in terms of concentrated density and dispersion.

Under the average impact threshold of a confounding variable framework, specifying the Δ- and Γ-relationships does not determine an exact *test*-statistic. Moreover, closed form estimation of its distribution is intractable. However, the distribution of confounding variable impact as well as the AITCV can be approximated by Monte Carlo simulation experiments. In particular, I develop a general method that first approximates the distribution of inference statistics resulting from the addition of a confounding variable to the BRM and then estimates its central tendency and dispersion. The approach I develop to assess the robustness of a causal inference supposes that strong ignorability is not satisfied in

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 Z \qquad (4.37)$$

but is satisfied in (4.10).

First, I approximate the average impact of an unobserved confounding variable and its distribution and subsequently find the threshold at the nominal α-level of 0.05. To identify the average impact, I use the weighted correlation of the link function (η) with the treatment and the confounding variable (4.22) to quantify the magnitude of the Γ- and Δ-relationships. In particular, I consider two different estimators of confounding variables in the parameter space. Both estimators employ the same basic approach in that they simulate hypothetical confounding variables based on weighted correlations. The first estimator, $\hat{\beta}_1^{(0)}$, utilizes the zero order weights. That is, those weights estimated in the original, unadjusted data which are based on the binomial variance only. For instance, I simulated hypothetical confounding variables based on correlations weighted by the inverse of the binomial variance:

$$n_j p_j (1 - p_j) \tag{4.38}$$

where n_j is the total sample size for covariate class j, and p_j is the probability of success within that covariate class. The second estimator, $\hat{\beta}_1^{(k)}$, utilizes the ML estimated weights resulting from estimation of (4.37). More specifically, this estimator uses weights from the final iteration of the IRWLS estimate from (4.37) to create hypothetical confounding variables. As a result, the weights in this estimator are adjusted to reflect the iterative solution to the model that only adjusts for the treatment (4.37). Using either weight, I simulate M confounding variables with specified Δ- and Γ-relationships and for each of the m confounding variables I estimate $\beta_{1_m}^*$ in

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1_m}^* Z + \beta_2 U_m \tag{4.39}$$

for m in 1 to M. Thus each $\beta_{1_m}^*$ represents the effect of the treatment, adjusting for the m^{th} hypothetical confounding variable of magnitude Δ and Γ. This process is repeated for each of the m simulated confounding variables thereby creating a distribution of possible coefficients, standard errors and *test*-statistics. I then defined the average impact of such a confounding variable as the mean inference statistic. To fully characterize and understand the impact of various confounders, I approximated the distribution of the inference statistics resulting from inclusion of the hypothetical confounder with specified relationships. In addition, I assessed the sensitivity of the approximated distribution to different study characteristics and model forms.

Example

Before addressing several complex issues surrounding AITCV's, I turn to a simple example to provide context. Similar to the example presented in Frank (2000), I apply the AITCV method to understand the relationship between family background and educational attainment but do so in an international context. The focus of this illustrative example is relationship between reading achievement and father's education level for sixth grade students in South Africa's Limpopo region. As indicated by Frank (2000) and previous studies in the United States, the relationship between family background and educational attainment has received a considerable amount of focus (e.g. Featherman & Hauser, 1976; Sobel, 1998). Similarly, within international contexts, the relationship of family background and achievement has been a consistent focus of inquiry (e.g. Buchmann, 2002; Lee, Zuze & Ross, 2005). In particular, as the educational level of a country's population embodies the human capital and human resources available for sustainable economic and social advancement, research that studies the influences, including family structures, that affect these educational levels are particularly relevant in building a country's infrastructure (Lee, Zuze & Ross, 2005). Further, understanding the roles of relevant family characteristics are of particular interest as such characteristics tend to have comparatively large influences on achievement (Buchman, 2002). As a result, international scholars have devoted much attention to "…improving knowledge of the ways that the family affects children's ability and motivation to learn and their academic achievement" (Buchmann, 2003, p. 4).

Differing to a certain extent from education in the U.S., a common characteristic of education throughout the developing world is grade repetition (Lee et al., 2005). Within sub-Saharan Africa repetition rates have been consistently higher than in other developing countries. For instance, repetition rates in primary grades in sub-Saharan Africa in 2000 were around 20% in grades one through five whereas corresponding rates in developed countries were about 1% and 3-10% in other developing countries such as Latin America (Table 6: Internal efficiency: Repetition in primary school, in: Nguyen., Wu, & Gillis, 2005). In sub-Saharan Africa eligibility for promotion to the next grade is heavily influenced by the student's ability to attain a minimum competency score on standardized tests. In particular, in the early grades, the literacy level of the student is of high relevance. In this descriptive example, I focus on the relationship between father's education level and a student's ability to meet the minimum competency score required. That is, our treatment or predictor of interest is the education level of a student's father and our outcome is a dichotomous response where one indicates the child has met the minimum requirements and zero indicates that he/she has not. In this example I used data from the United Nations' International Institute for Educational Planning (IIEP) and the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) programs (Ross, Saito, Dolata, Ikeda, & Zuze 2004). Further, father's education level is an ordinal variable constructed by researchers in SACMEQ where one indicate no formal schooling, two indicates some primary schooling, three indicates the completion of primary school, four indicates some secondary schooling, five indicates the completion of secondary school and six, indicates any exposure to post-secondary education. For illustrative purposes, I first strictly focus on unconditional comparisons or in other words

I do not control for any other variables thought to be relevant to educational attainment. Though father's education level likely contributes to his child's educational attainment, it is likely that such estimates are inflated as such an analysis fails to take into account other factors such as mother's education level. The question of interest here focuses on the inference of whether father's education would continue to have a significant association even when controlling for other factors. That is, given that there is a significant association between father's education and a student's attainment, how large must the relationships between a confounding variable and outcome (Δ-relationship) as well as the confounding variable and the treatment (Γ-relationship) be to invalidate the inference that father's education level is significantly associated with educational attainment?

To address this question I utilized the AITCV framework described above. Using the AITCV method and simulation I looked to estimate the magnitude of the Δ- and Γ-relationships of a confounding variable that would reduce the original *test*-statistic to an average *test*-statistic of 1.96. More specifically, I first estimated the treatment effect, β_1, in (4.1) using the SACMEQ data and the respective BRM (Table (4.40)). My unconditional estimates indicated that there is a significant positive association between attaining the minimum literacy level and father's education level. That is, the number of students who meet the minimum literacy level tends to increase as father's education increases. As such associational inferences did not attempt to adjust for any confounding variables, I then asked what the magnitude of the Δ- and Γ-relationships for a confounding variable would have to be to change the original significant inference into an insignificant finding. To assess the robustness of the original inference using the AITCV framework I then estimated the weighted correlation between the logit of success and father's education. In this artificial example that utilizes no controls, the unweighted/weighted correlation between educational attainment and father's education level is around 0.91/0.92 indicating a strong linear relationship. The strength of such unconditional relationships in developing countries is often provided as evidence of the vast inequalities in such countries (e.g. Baker, Goesling & Letendre, 2002; Lee et al. 2005).

Table(4.40): Regression of Minimum Educational Competency on Father's Education Level

	Coefficient	Standard Error	Z-value
Intercept (β_0)	-3.01	0.38	-7.95
Father Education (β_1)	0.37	0.09	4.38

Next, I simulated multiple sets of confounding variables to identify the AITCV. More specifically, I first simulated a confounding variable with Δ- and Γ-relationships equal to 0.75 and re-estimated the BRM controlling for the simulated confounding variable. Next, I repeated this process 1000 times and saved each new *test*-statistic for father's education controlling for the confounding variable each time. I then averaged the 1000 *test*-statistics to identify the average impact of a confounding variable with Δ- and Γ-relationships equal to 0.75. Given this data my simulations indicated that a confounding variable with a 0.75 Γ- and Δ-relationship would, on average, reduce the *test*-statistic from 4.38 to 2.6. As a result, I repeated this process with increasing Γ- and Δ-relationships until the average *test*-statistic was approximately 1.96. My analyses for this illustrative example indicated that the AITCV would be approximately 0.69 indicating that the Δ- and Γ-relationships would need to be approximately 0.83. In other words, in order to alter our original inference we would need a confounding variable or set of confounding variables with Δ- and Γ-relationships roughly equal to 0.83. Though such relationships seem excessively large, unconditional correlations between socio-economic status proxies such as father's education and academic attainment in developing worlds tend to be strong (e.g. Baker, Goesling &Letendre, 2002). Although, I do not speak to the likelihood of a confounding variable exceeding such relationships, it does quantify the magnitude of relationships needed by confounding variable to indicate that father's education is not significantly associated with educational attainment. Further such likelihood extensions can be addressed using Pan and Frank (2004) or Frank (2000).

Estimating the AITCV

Another difficulty resulting from the iterative nature of estimation is the complex interplay between the Γ- and Δ-relationships. Unlike the *t*-statistic resulting from OLS inference, the *test*-statistic resulting from MLE inference is not necessarily minimized when $\Delta = \Gamma$ ($\rho_{\eta x} = \rho_{\eta u}$). As a result, estimation of an impact threshold requires numerical estimation for each data set. To develop such estimation techniques I first established the monotonicity of the average *test*-statistic given the impacts through simulation. That is,

$$E[z(\widehat{\delta^*}) \mid k^{(ij)}] > E[z(\widehat{\delta^*}) \mid k^{(lm)}] \tag{4.41}$$

where

$$k^{(ij)} = \rho^{(i)}_{UZ}\rho^{(j)}_{U\eta} \quad \text{and} \quad k^{(ij)} < k^{(lm)} \tag{4.42}$$

Using this property, I developed an iterative method that estimates the AITCV for a given data set. Initially the method sets the AITCV equal to the WLS ITCV. Next it evaluates the average inference statistic given the WLS ITCV. It then adjusts the AITCV up or down based on the previous evaluation and continues iteratively until the adjustment is sufficiently small.

Extending the AITCV to Multiple Predictors

Though the previous approaches have defined the ITCV for BRMs and outlined a practical method to estimate it, its use is limited to models that consider only a treatment and a single confounder. In order to make this method practically useful in the social sciences, we must extend it to include multiple covariates. For instance, we may estimate a treatment effect controlling for a number of known and measured covariates and then be interested in assessing the estimate's sensitivity to an unmeasured confounder. That is, we are interested in

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 Z + \beta_2 X_1 + \beta_3 X_2 + \ldots + \beta_{p+1} X_p + \beta_{p+2} U \tag{4.43}$$

Such a model supposes that the outcome is influenced by multiple factors[2]. Drawing on Frank (2000), I redefine the relevant correlations as partial correlations which control for ***X*** where ***X*** is the vector consisting of covariates $\{X_1,\ldots X_p\}$. In other words, we replace the correlations in (4.26) with the following:

$$r_{\eta \mathrm{z}} \text{ becomes } r_{\eta \mathrm{z}|\mathbf{x}} = \frac{r_{\eta \mathrm{z}} - r_{\eta \mathbf{x}} r_{\mathbf{x}\mathrm{z}}}{\sqrt{(1-r_{\eta \mathbf{x}}^2)(1-r_{\mathbf{x}\mathrm{z}}^2)}} \tag{4.44}$$

$$r_{\mathrm{zu}} \text{ becomes } r_{\mathrm{zu}|\mathbf{x}} = \frac{r_{\mathrm{zu}} - r_{\mathrm{z}\mathbf{x}} r_{\mathbf{x}\mathrm{u}}}{\sqrt{(1-r_{\mathrm{z}\mathbf{x}}^2)(1-r_{\mathbf{x}\mathrm{u}}^2)}} \tag{4.45}$$

$$r_{\eta \mathrm{u}} \text{ becomes } r_{\eta \mathrm{u}|\mathbf{x}} = \frac{r_{\eta \mathrm{u}} - r_{\eta \mathbf{x}} r_{\mathbf{x}\mathrm{u}}}{\sqrt{(1-r_{\eta \mathbf{x}}^2)(1-r_{\mathbf{x}\mathrm{u}}^2)}} \tag{4.46}$$

where

$$r_{\eta \mathbf{x}}^2 = \sum_{i=1}^{n} \gamma_i r_{\eta x_i} \qquad r_{\mathbf{x}\mathrm{z}}^2 = \sum_{i=1}^{n} \pi_i r_{z x_i} \qquad r_{\mathbf{x}\mathrm{u}}^2 = \sum_{i=1}^{n} \lambda_i r_{u x_i} \tag{4.47}$$

and γ_i, π_i and λ_i are obtained by regressing η, *Z* and *U* on the covariates ***X***. As a result an estimator corresponding to (4.26) which controls for multiple covariates can be written as

$$\beta_1^{wls} = \left(\frac{s_\eta}{s_z}\right)\left(\frac{r_{\eta z|\mathbf{x}} - r_{\eta u|\mathbf{x}} r_{zu|\mathbf{x}}}{1 - r_{zu|\mathbf{x}}^2}\right) \tag{4.48}$$

where each correlation and standard deviation in (4.44) to (4.48) is appropriately weighted. Accordingly, our speculation parameters now become

$$\Gamma = r_{zu|\mathbf{x}} \text{ and } \Delta = r_{\eta u|\mathbf{x}} \tag{4.49}$$

where the corresponding zero order correlations are determined by the equations

$$r_{zu} = r_{zu|\mathbf{x}}\sqrt{(1-r_{z\mathbf{x}}^2)(1-r_{\mathbf{x}u}^2)} + r_{z\mathbf{x}} r_{\mathbf{x}u} \tag{4.50}$$

$$r_{\eta u} = r_{\eta u|\mathbf{x}}\sqrt{(1-r_{\eta\mathbf{x}}^2)(1-r_{\mathbf{x}u}^2)} + r_{\eta\mathbf{x}} r_{\mathbf{x}u} \tag{4.51}$$

Departing from Frank (2000), I do not assume the weighted correlation between the covariates, ***X***, and the unobserved confounding variable, *U*, to be zero. That is, though non-zero relationships between the unobserved variable, *U,* and the observed variables, ***X***, partially absorb the impact of *U* on the estimated treatment effect in the linear model (Frank, Maurolis, Duong & Kelcey, 2009), such relationships do not necessarily absorb *U*'s impact in BRMs. Thus to consider the maximum impact of an unmeasured confounder in BRMs, one can not constrain such relationships to be zero. As a result, I allowed the correlations between the covariates and the unobserved confounding variable to vary by sequentially drawing random values from the Uniform distribution. More specifically, the bounds of the Uniform distribution were deflated from the interval [-1, 1] to recognize the constraints of several concurrent correlations. That is, the correlation between any two variables is constrained by the relation that each variable has with all other variables. In particular, this constraint is

$$\rho_{x_1x_2} \in \rho_{ux_1}\rho_{ux_2} \pm \sqrt{(1-\rho_{ux_1}^2)+(1-\rho_{ux_2}^2)} \tag{4.52}$$

As a result, I estimated the constraints on the correlation between x_1 and the confounding variable, U using (4.52). Using the constrained range as bounds, I drew a random correlation value from U($\rho_{ux_1}^{(lower)}$, $\rho_{ux_1}^{(upper)}$) to represent the correlation between x_1 and the confounding variable, U. In a similar manner, I estimated the bounds for the correlation between x_2 and the confounding variable, U and randomly drew another value from the Uniform distribution to represent their correlation. This process was repeated until each correlation was randomly selected.

Subsequently, I assessed whether the distribution of the resulting *test*-statistics is well approximated by the same distribution as in the simple regression case. I further assessed the sensitivity of the results to variation in several parameters. In particular, I examined how both the AITCV and the approximate distribution of the impacts would change as I altered the probability of success from its default of 0.5 to 0.1 and 0.9 and sample size (from *n*=50 to *n*=25 and *n*=500). In addition, I examined the sensitivity of the results to the assumed unconditional distribution of the unobserved confounder. That is, I assessed how the results might change when assuming *U* comes from a Beta rather than Uniform distribution.

Finally, I examined the sensitivity of such results to the choice of the link function. First, I examined if the *test*-statistics resulting from models using a probit link followed a similar distribution. Second, I assessed the sensitivity of the link specification by allowing the true link function to be a probit and estimating both the model and the AITCV with the logit link. Such sensitivities help to understand the robustness of using an AITCV approach.

Results of AITCV Estimation

By varying a variety of data generation parameters, I simulated binomial outcomes, covariates, a treatment and a confounding variable. Through this simulation, I explored four dimensions of the AITCV framework for BRMs. I first examined the average impact of a confounding variable in simple BRMs. That is, initially, I examined how a confounding variable with various relationships affects the *test*-statistic of the treatment when there are no other covariates. In other words, I did not attempt to estimate the threshold but rather, first, understand how confounding variables of different magnitudes change the *test*-statistic of the treatment effect. In a similar manner, I next assessed the AITCV in BRMs with multiple covariates. For example, in assessing the effect of a treatment when there are measured confounders one includes those measured confounders in the regression model to better estimate the treatment effect. Third, I explored the sensitivity of such results to variations in several relevant parameters. Finally, I assessed the AITCV method and compared such thresholds to that of the original ITCV framework.

Within each of these dimensions, I focused my inquiry on four aspects of the framework. First, I assessed the accuracy of the AITCV method in estimating the true *test*-statistic. In other words, I estimated the nominal coverage of the AITCV. Second, I approximated the distribution of such *test*-statistics corresponding to β_1^* in (4.39). Third, I examined the monotonicity of the average *test*-statistic given impacts. That is, as the product, k, of the speculation parameters (Γ and Δ) increases, the average *test*-statistic decreases. Finally, I assessed the sensitivity of the impact on the *test*-statistic to the relative contribution of Γ and Δ to their product k. That is, for a given k how does the impact on the *test*-statistic change when, for example, $\Gamma >> \Delta$ (e.g. $\Gamma \approx k$ and $\Delta \approx 1$)?

Simple Regression-Coverage of AITCV

First, I addressed the impact of a confounding variable in BRMs with no covariates beyond the treatment as shown in equation (4.37). In particular, given a treatment effect and corresponding *test*-statistic, I asked how the treatment's *test*-statistic resulting from equation (4.37) would change with the inclusion of an unmeasured variable, *U,* with specified Δ- and Γ-relationships as in equation (4.39). To assess this I first randomly generated an unmeasured confounding variable from the Uniform distribution within the bounds of -5 to 5 with said relationships and recorded the new *test*-statistic of the treatment when controlling for this unmeasured confounding variable. Subsequently, I withheld the unmeasured confounding variable from estimation and generated 100 proxy confounding variables with the same Δ- and Γ-relationships from the same Uniform distribution. I reran (4.39) using the each proxy confounding variable (one at a time) and recorded the corresponding *test*-statistics of the treatment. The 100 *test*-statistic created a distribution of possible *test*-statistics for the treatment effect when controlling for an unmeasured confounding variable with specified Δ- and Γ-relationships. To understand properties of this approach I repeated this process 100 times for multiple combinations of Δ, Γ and Φ displayed subsequently.

The results of the experiment suggested that the simulation method outlined above is accurate among a wide array of covariance structures. Table (4.53) provides a summary for several combinations of the Φ-, Γ-, Δ- relationships. This table includes the coverage of the true *test*-statistic (i.e. the treatment's *test*-statistic when controlling for the true confounding variable) by the *test*-statistics corresponding to controlling for the simulated confounding variables. Further, to summarize the range and dispersion of the *test*-statistics corresponding to controlling for the simulated confounding variables, the table includes the average *test*-statistic and the standard deviations of those simulated *test*-statistics. In particular the first three columns specify the weighted partial correlations between the logit of the outcome and the treatment, Φ, between the treatment and the unobserved confounder, Γ, and between the logit of the outcome and the unobserved confounder, Δ. The next set of columns describes the coverage of the true *test*-statistic by the simulated *test*-statistic distribution using the standard deviations of the distribution. The next column then describes the average size of the corresponding standard deviations. Finally, the last column displays the percent of times one can reject the null hypothesis that the distribution of potential *test*-statistics resulting from the adjustment for an unmeasured confounding variable with Γ- and Δ-relationships is a Beta distribution.

Using the AITCV framework the estimated average *test*-statistic fell within ± 0.1 of the true *test*-statistic 52% of the time and fell within ± 0.35 of the true *test*-statistic 96% of the time. That is, the distribution of $z(\beta_1^*)$ from (4.39) tends to be tightly concentrated and the true *test*-statistic, $z(\beta_1)$, of the corresponding maximum likelihood estimate of the treatment effect, β_1, in (4.10) tends to be close to the center of the distribution. Similarly, approximately 83% of the time the true *test*-statistic fell within plus or minus two standard deviations of the average impact of a confounding variable with said relationships. Further, the simulations tended to indicate that such standard deviations tended to grow as the impact grew. That is, as the product of the Γ- and Δ-relationship grew, the distribution of possible *test*-statistics became more dispersed. Finding an increasing imbalance between treatment groups to be associated with an increasing range of possible inference statistics is well aligned with prior research (e.g. Rosenbaum, 1995). However, whereas in other approaches the plausible inference statistic tends to grow unabated and in an exponential manner as the imbalance grows, in the current approach the range grows slowly and such growth virtually discontinues. Further, the estimates in Table (4.53) are based on updated weights (4.37) as the zero order weights (4.38) had the same general results but produced approximately 30% wider intervals.

Table(4.53): Examples of coverage of true test-statistic by estimated test-statistic distribution (in percent) and goodness of fit test for the empirical distribution vs. beta distribution

Correlations			Coverage (proportion) in ± Standard Deviations			Ave. SD	χ^2 G.o.F. Beta[1] Rejected (%)
Φ	Γ	Δ	1	2	3		
0.30	0.10	0.10	0.54	0.82	0.96	0.029	9
		0.50	0.52	0.78	0.94	0.085	1
		0.70	0.57	0.87	0.94	0.143	7
	0.50	0.10	0.59	0.82	0.97	0.138	6
		0.50	0.55	0.85	0.98	0.137	6
		0.70	0.47	0.84	0.93	0.146	7
	0.70	0.10	0.5	0.79	0.96	0.192	23
		0.50	0.53	0.77	0.96	0.177	14
		0.70	0.48	0.79	0.95	0.133	3
0.50	0.10	0.10	0.51	0.87	0.99	0.028	5
		0.50	0.5	0.8	0.95	0.086	5
		0.70	0.51	0.8	0.94	0.122	9
	0.51	0.10	0.47	0.83	0.96	0.117	1
		0.50	0.43	0.79	0.92	0.125	4
		0.70	0.58	0.88	0.99	0.137	9
	0.70	0.10	0.5	0.9	0.97	0.156	8
		0.50	0.53	0.85	0.97	0.176	11
		0.70	0.45	0.85	1.00	0.150	9
0.70	0.10	0.10	0.61	0.87	0.97	0.021	13
		0.50	0.54	0.83	0.97	0.100	4
		0.70	0.59	0.84	0.98	0.094	3
	0.50	0.10	0.58	0.89	0.98	0.089	3
		0.50	0.5	0.85	0.98	0.103	4
		0.70	0.56	0.85	0.98	0.127	6
	0.70	0.10	0.48	0.81	0.97	0.116	4
		0.50	0.53	0.86	0.98	0.124	5
		0.70	0.47	0.83	0.96	0.137	6
	Average					0.118	6.8

[1] Number of times H_0 is rejected out of 100 (H_0: Distribution of test-values is Beta(α,β), where α, β are estimated via MoM & ML)

The second result suggested the distribution of these *test*-statistics is well approximated by a Beta distribution. More specifically I compared the simulated distribution of the resulting *test*-statistics to a theoretical Beta distribution in three ways. To compare them, first I used Pearson's *chi*-squared goodness of fit test, second I contrasted their respective moments and third I examined the quantile-quantile plots. In conducting the *chi*-squared goodness of fit test, I first estimated the parameters of the Beta distribution as its location depends on the given dataset. To do this, I first estimated the parameters by using the method of moments estimator. Next, using the estimates produced by the method of moments as starting values, I estimated the parameters using the maximum likelihood estimator. Using these parameter estimates, I calculated the theoretical density of the beta distribution. I then divide the data into bins based on a formula that asymptotically minimizes the integrated mean squared error (Scott, 1979) and conducted a Pearson *chi*-squared goodness of fit distribution test based on the null hypothesis that the empirical distribution comes from a beta distribution with the estimated parameters. This statistic

$$\mathrm{X}^2 = \sum_k \left(\frac{(O-E)^2}{E} \right) \tag{4.54}$$

has an asymptotic *chi*-squared distribution with *K*-2-1 degrees of freedom (where *K* is the number of bins, less 2 since estimating the two parameters of the Beta distribution). The result of the test are presented in the final column of Table (4.53). The results indicated that the *chi*-squared goodness of fit test was able to reject the null hypothesis at the nominal α-level of 0.05 an average of 6.8% of the time throughout a variety of data structures. I speculated that the additional 1.8% may be the result of two factors. First, the distribution of the test statistic is only χ^2 as sample size goes to infinity. Second, there is a considerable potential for small bin counts (e.g. <5) when estimating empirical distributions estimated by only 100 data points.

In contrasting the moments of the theoretical Beta distributions with the moments of the simulated *test*-statistic distribution, I compared the first four moments. Again the parameters of the theoretical are estimated by ML with the method of moments estimators providing the starting value. Table (4.55) contrasts the observed empirical moments with the expected moments. We see that the empirical distribution corresponds highly with the first two moments of the respective theoretical Beta distribution. However, in comparing the third and fourth moments, skew and kurtosis, some departure was evident.

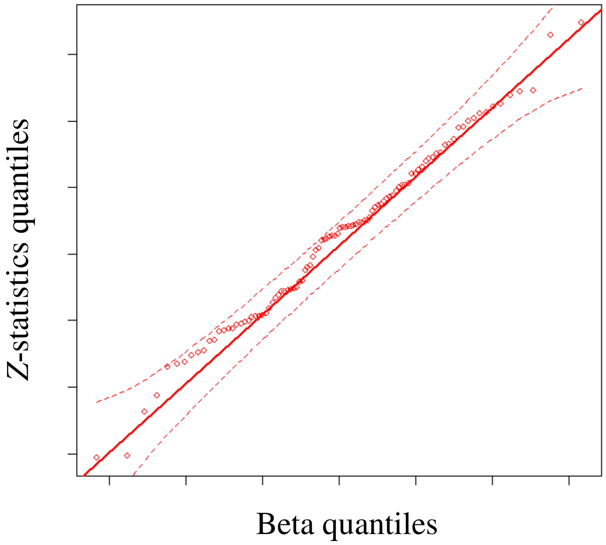
Table(4.55): Comparison of first four moments of the distribution of the *test*-statistics estimated by the AITCV method and the corresponding Beta distribution

Φ	Γ	Δ	Expected Mean	Observed Mean	Expected Variance	Observed Variance	Expected Skew	Observed Skew	Expected Kurtosis	Observed Kurtosis
0.30	0.10	0.10	0.50	0.50	0.00	0.00	0.00	-0.09	-0.03	-0.16
		0.50	0.50	0.50	0.01	0.01	0.00	-0.05	-0.16	-0.07
		0.70	0.50	0.50	0.02	0.02	0.00	-0.04	-0.44	-0.04
	0.50	0.10	0.50	0.50	0.02	0.02	0.00	-0.01	-0.41	-0.16
		0.50	0.50	0.50	0.02	0.02	0.00	-0.06	-0.40	-0.11
		0.70	0.50	0.50	0.02	0.02	0.00	-0.13	-0.46	-0.09
	0.70	0.10	0.50	0.50	0.04	0.04	0.00	0.01	-0.73	-0.12
		0.50	0.50	0.50	0.03	0.03	0.00	-0.02	-0.63	-0.21
		0.70	0.50	0.50	0.02	0.02	0.00	-0.17	-0.38	0.05
0.50	0.10	0.10	0.50	0.50	0.00	0.00	0.00	-0.13	-0.02	-0.15
		0.50	0.50	0.50	0.01	0.01	0.00	0.02	-0.17	-0.11
		0.70	0.50	0.50	0.02	0.02	0.00	-0.03	-0.33	-0.05
	0.51	0.10	0.50	0.50	0.01	0.01	0.00	-0.01	-0.31	-0.20
		0.50	0.50	0.50	0.02	0.02	0.00	0.04	-0.34	-0.23
		0.70	0.50	0.50	0.02	0.02	0.00	-0.05	-0.41	-0.02
	0.70	0.10	0.50	0.50	0.03	0.03	0.00	-0.03	-0.51	-0.15
		0.50	0.50	0.50	0.03	0.03	0.00	-0.04	-0.62	-0.22
		0.70	0.50	0.50	0.02	0.02	0.00	-0.04	-0.47	-0.08

0.70	0.10	0.10	0.50	0.50	0.00	0.00	0.00	-0.29	0.00	0.02
		0.50	0.50	0.50	0.01	0.01	0.00	0.14	-0.23	-0.16
		0.70	0.50	0.50	0.01	0.01	0.00	0.08	-0.20	-0.20
	0.50	0.10	0.50	0.50	0.01	0.01	0.00	-0.07	-0.18	-0.17
		0.50	0.50	0.50	0.01	0.01	0.00	0.00	-0.24	-0.19
		0.70	0.50	0.50	0.02	0.02	0.00	0.07	-0.35	0.00
	0.70	0.10	0.50	0.50	0.01	0.01	0.00	-0.16	-0.29	-0.13
		0.50	0.50	0.50	0.02	0.02	0.00	-0.02	-0.34	-0.18
		0.70	0.50	0.50	0.02	0.02	0.00	-0.02	-0.40	-0.17

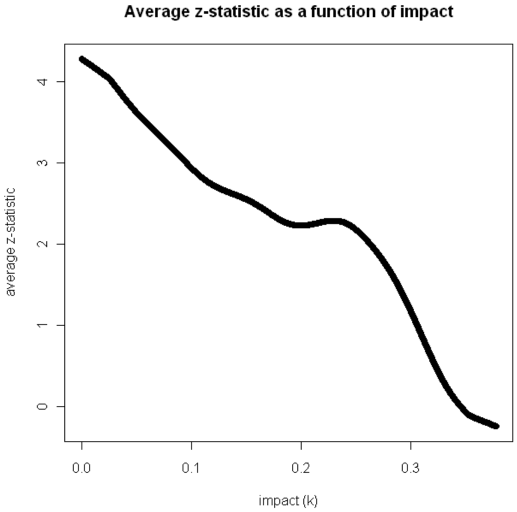
Table (4.55) suggests that the first two moments of these distributions tend to be identical whereas the third moment tends to deviate slightly (generally correct to the first decimal point) though in a symmetric manner. The fourth moment, however, tends to deviate to a higher degree. In addition to numerical assessment, I produced a quantile-quantile plots to graphically compare the quantiles of the generated distribution for each dataset with that of the Beta distribution. Figure (4.56) presents a typical quantile-quantile plot for such data.

Figure(4.56): Example of typical quantile-quantile plot for distribution of *test*-values vs. beta distribution for simple regression



Next, such simulations provided evidence that on average as one increases the impact or product, *k,* of the Γ- and Δ-relationships, the average *test*-statistic tends to decrease (Figure (4.57)). More specifically, through a wide range of relationships, the simulations indicated that the mean *test*-statistic produced for an impact decreases as the impact increases. Such results suggest that the average *test*-statistic has an inverse monotonic relationship with the impact which may facilitate an iterative threshold search procedure. That is, given the average *test*-statistic decreases as the impact increases, one can search for the AITCV using an algorithm that first estimates the threshold using WLS and then adjusts the threshold upwards (increase magnitude of impact) if the WLS threshold is above the nominal significance level (e.g. 1.96) and downwards (decrease the magnitude of the impact) if the WLS threshold is below the nominal significance level.

Figure(4.57): Example of relationship between the average *test*-statistic and the impact (product of Γ- and Δ-relationships) for simple regression



Finally, I assessed how the split of *k* influences the average *test*-statistic. For example, such an assessment asks if the AITCV is *k*=0.25, how does the AITCV change when Γ=0.5 while Δ=0.5 as opposed to Γ=0.3125 while Δ=0.8 or similarly when Γ=0.8 while Δ=0.3125? The simulations indicated that the split of the impact, *k,* has relatively little influence over the average treatment effect *test*-statistic resulting from the inclusion of a confounding variable. More specifically, though the average *test*-statistic is highly dependent on *k*, when holding *k* constant, the split of the product between the Γ- and Δ-relationships generally only changes the second decimal point of the average *test*-statistic. Such results are revisited subsequently in the multiple regression case.

Multiple Regression-Coverage of AITCV

In a manner similar to above, I assessed the impact of a confounding variable in BRMs with other measured covariates. That is, suppose an incomplete model of an outcome is

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 Z + \beta_2 X_1 + \beta_3 X_2 + \ldots + \beta_{p+1} X_p \tag{4.58}$$

as it excludes an important yet unmeasured confounding variable, *U*. As a result the correct model is actually

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 Z + \beta_2 X_1 + \beta_3 X_2 + \ldots + \beta_{p+1} X_p + \beta_{p+2} U \tag{4.59}$$

In particular, given a treatment effect and corresponding *test*-statistic that was estimated in the presence of nine other covariates e.g. (4.58), I now asked how the treatment's *test*-statistic would change with the inclusion of an unmeasured variable (e.g. equation (4.59)) with specified Δ- and Γ-relationships where these relationships are now specified as partial correlations (4.49). To assess this I first randomly generated an unmeasured confounding variable with said relationships from the Uniform distribution with bounds negative 5 to 5 and recorded the new *test*-statistic of the treatment when controlling for measured variables and unmeasured confounding variable. Subsequently, I withheld the unmeasured confounding variable from estimation and generated 100 proxy confounding variables with the same Δ- and Γ-relationships from the same uniform distribution. This created a distribution of possible *test*-statistics for the treatment effect when controlling for an unmeasured confounding variable with specified Δ- and Γ-relationships. This process was repeated 100 times for multiple combinations of Δ, Γ and Φ.

The results of the experiment with multiple confounding variables suggested that the AITCV method improves as the number of covariates increases. Table (4.60) provides a summary for several combinations of the Φ-, Γ-, Δ- relationships, their coverage and their average standard deviations. Using the AITCV framework for multiple regression the estimated average *test*-statistic fell within ± 0.06 of the true *test*-statistic 68% of the time and fell within ± 0.12 of the true *test*-statistic 96% of the time. That is, the distribution of the $z(\beta_1^*)$'s tended to be even more tightly concentrated around the true *test*-statistic, $z(\beta_1)$, when compared with that of the simple regression model above.

Table(4.60): Examples of coverage of true test-statistic by estimated test-statistic distribution (in percent) and goodness of fit test for the empirical distribution vs. beta distribution when there are multiple measured confounders

Correlations			Coverage (proportion) in ± Standard deviations			Ave. SD	χ^2 G.o.F. Beta[1] Rejected (%)
Φ	Γ	Δ	1	2	3		
0.30	0.10	0.10	0.677	0.953	0.997	0.021	6
		0.50	0.664	0.952	0.999	0.034	3
		0.70	0.687	0.952	0.998	0.038	5
	0.50	0.10	0.693	0.950	0.996	0.046	4
		0.50	0.678	0.958	0.996	0.046	6
		0.70	0.676	0.954	0.998	0.048	7
	0.70	0.10	0.674	0.950	1.000	0.090	10
		0.50	0.666	0.960	1.000	0.086	8
		0.70	0.685	0.948	0.997	0.099	3
0.50	0.10	0.10	0.690	0.946	1.000	0.010	5
		0.50	0.701	0.950	0.997	0.022	5
		0.70	0.684	0.957	0.995	0.057	6
	0.50	0.10	0.657	0.963	0.999	0.050	1
		0.50	0.677	0.953	1.000	0.080	4
		0.70	0.665	0.959	1.000	0.065	7
	0.70	0.10	0.689	0.956	0.997	0.065	6
		0.50	0.658	0.960	1.000	0.095	18
		0.70	0.675	0.955	0.997	0.056	6
0.70	0.10	0.10	0.671	0.952	0.998	0.013	9

	0.50	0.696	0.950	0.997	0.028	4
	0.70	0.658	0.959	0.998	0.063	3
0.50	0.10	0.657	0.963	0.999	0.050	3
	0.50	0.683	0.953	0.997	0.045	4
	0.70	0.669	0.956	0.999	0.069	6
0.70	0.10	0.661	0.964	0.999	0.118	4
	0.50	0.662	0.962	1.000	0.092	5
	0.70	0.691	0.958	0.995	0.067	6
Average		0.676	0.955	0.998	0.057	5.7

[1] Number of times H_0 is rejected out of 100 (H_0: Distribution of test-values is Beta(α,β), where α, β are estimated via MoM & ML)

In addition, similar to the simple regression model results, as the impact grew so did the dispersion of the potential *test*-statistics. However, when there are multiple covariates in the model, I tended to see a diminished increase in the dispersion as the impact increased. In other words, not only is the impact of an unobserved confounding variable partially absorbed by measured covariates, but also the dispersion of the impact is partially absorbed. That is, the inter-relationships among the measured variables and the unmeasured variable allow the measured variables to act as proxies for the unmeasured variable. As a result, improving the balance between treatment groups or the quality of the controls in the model via measured variables absorbs the impact of an unmeasured confounding variable and narrows the range of the effect that impact can have on the *test*-statistic.

Such results are also consistent with the simple regression model above in that the AITCV method helps us better understand the effect of an unmeasured confounding variable with much more precision than other methods. As the interval of possible inference statistics grows as the imbalance grows in canonical methods, the AITCV tends to restrict the growth of such intervals. As the AITCV parameterizes the Δ- and Γ-relationships as partial correlations, it is evident that the minimal growth in the range of possible inference statistics associated with increased imbalance between groups is not directly due to the absorption of the unobserved confounding variable by measured variables. Rather, including multiple variables to some extent stabilizes the Fisher scoring iterations so that the adjusted dependent variable tends to change relatively little in comparing models (4.58) and (4.59). As a result, though an unobserved confounding variable with relatively large Γ- and Δ-relationships can have a large impact on the treatment effect estimate and its corresponding *test*-statistic, the range of its possible impact is restricted. For example, for a given dataset with a treatment that has a partial correlation with the logit of the outcome of 0.3 controlling for the measured covariates, the impact of an unmeasured confounding variable with a Γ-relationship of 0.1 and a Δ-relationship of 0.1 on the treatment's original *test*-statistic can be identified within ± 0.042 of the true *test*-statistic over 95% of the time. Now if one asks a similar question but replaces the Γ- and Δ-relationships with higher partial correlation of say 0.7, the interval of possible *test*-statistics widens from ±0.042 to ±0.099 but coverage stays nearly the same at approximately 95%. In other words, for a given dataset, the AITCV framework would indicate that if you included a variable with Γ=Δ=0.1 relationships, your *test*-statistic would drop from say 3.5 to say 3.1 ± 2(0.042) 95 % of the time. Similarly, if you included a variable with Γ=Δ=0.7 relationships your *test*-statistic would drop from say 3.5 to say 1.5 ± 2(0.099) 95% of the time.

Next, corresponding with the simple regression case, the results from the multiple regression models suggested that the distribution of *test*-statistics is well approximated by a Beta distribution. I compared the simulated distribution of the resulting *test*-statistics to a theoretical beta distribution using Pearson's *chi*-squared goodness of fit test, respective moments and quantile-quantile plots. The results of the *chi*-squared test are presented in the final column of Table (4.60). The results indicated that the *chi*-squared goodness of fit test was able to reject the null hypothesis at the nominal α-level of 0.05 an average of 5.7% of the time throughout a variety of data structures.

I next compared the first four moments of the theoretical Beta distribution with the moments of the simulated *test*-statistics distribution. Table (4.61) presents the results of the moments. Similar to the simple regression case, but more closely aligned, we see a high correspondence between the empirical distribution and the respective theoretical Beta distribution for the first two moments. When comparing the third and fourth moments, skew and kurtosis, we again saw some departure, however the departure was less than that of the simple regression case.

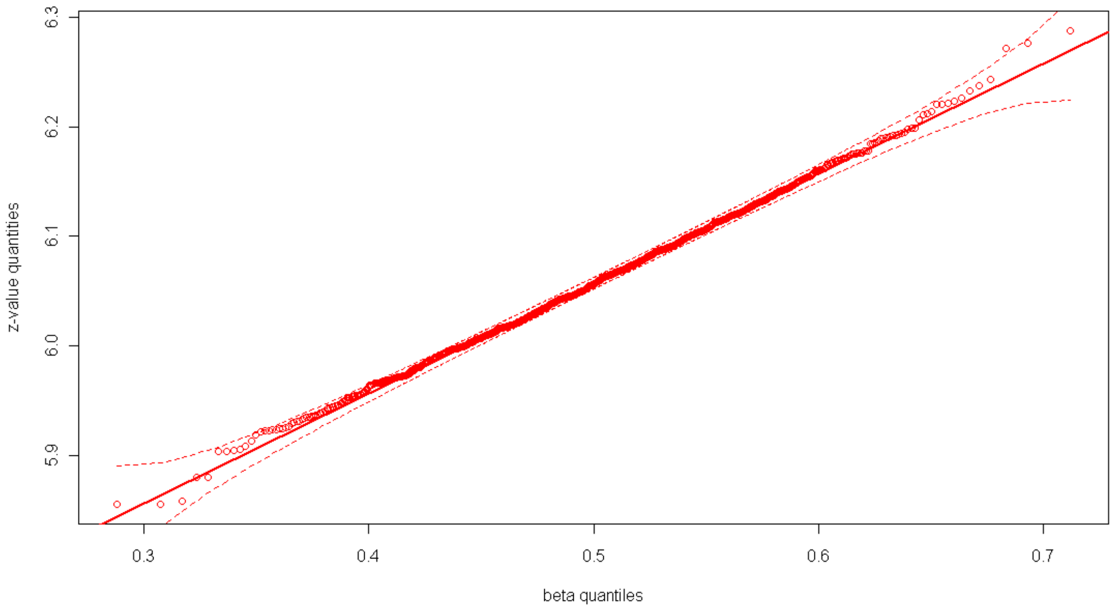
Table(4.61): Comparison of first four moments of the distribution of the *test*-statistics estimated by the AITCV method and the corresponding Beta distribution

Φ	Γ	Δ	Expected Mean	Observed Mean	Expected Variance	Observed Variance	Expected Skew	Observed Skew	Expected Kurtosis	Observed Kurtosis
0.30	0.10	0.10	0.50	0.50	0.00	0.00	-0.04	-0.12	-0.08	-0.05
		0.50	0.50	0.50	0.00	0.00	0.04	-0.13	0.06	0.01
		0.70	0.50	0.50	0.00	0.00	0.03	-0.27	-0.02	0.22
	0.50	0.10	0.50	0.50	0.00	0.00	0.00	-0.11	-0.03	0.06
		0.50	0.50	0.50	0.00	0.00	0.01	0.11	-0.04	0.14
		0.70	0.50	0.50	0.00	0.00	0.01	0.08	-0.02	0.04
	0.70	0.10	0.50	0.50	0.01	0.01	-0.01	0.08	-0.20	-0.18
		0.50	0.50	0.50	0.01	0.01	-0.01	0.03	-0.22	-0.41
		0.70	0.50	0.50	0.01	0.01	0.00	0.00	-0.17	-0.07
0.50	0.10	0.10	0.50	0.50	0.00	0.00	0.02	-0.14	-0.02	-0.10
		0.50	0.50	0.50	0.00	0.00	-0.03	0.02	0.04	0.03
		0.70	0.50	0.50	0.00	0.00	-0.01	-0.20	-0.08	0.03
	0.51	0.10	0.50	0.50	0.00	0.00	0.00	-0.03	-0.03	-0.26
		0.50	0.50	0.50	0.01	0.01	-0.02	-0.01	-0.19	-0.34
		0.70	0.50	0.50	0.00	0.00	-0.01	-0.04	-0.07	-0.28
	0.70	0.10	0.50	0.50	0.00	0.00	0.01	0.08	-0.18	0.03
		0.50	0.50	0.50	0.01	0.01	0.00	-0.01	-0.23	-0.36
		0.70	0.50	0.50	0.00	0.00	0.02	-0.01	-0.01	-0.17
0.70	0.10	0.10	0.50	0.50	0.00	0.00	-0.02	-0.18	0.01	-0.15
		0.50	0.50	0.50	0.00	0.00	-0.01	-0.01	-0.01	0.15
		0.70	0.50	0.50	0.00	0.00	0.02	0.09	-0.06	-0.24
	0.50	0.10	0.50	0.50	0.00	0.00	0.00	0.08	-0.02	-0.01

	0.50	0.50	0.50	0.00	0.00	0.00	0.21	-0.04	-0.01
	0.70	0.50	0.50	0.00	0.00	-0.01	-0.07	-0.16	-0.27
0.70	0.10	0.50	0.50	0.01	0.01	0.03	-0.05	-0.28	-0.46
	0.50	0.50	0.50	0.01	0.01	0.00	0.01	-0.22	-0.44
	0.70	0.50	0.50	0.00	0.00	0.01	0.11	-0.14	0.00

To graphically assess the fit of the Beta distribution, I also produced quantile-quantile plots to compare the distributions. Similar to that of the simple regression case, the plots indicated a close alignment. However, such Q-Q plots in the multiple regression case tend to represent much more concentrated Beta distributions than their simple regression analogues. Figure (4.62) presents a typical quantile-quantile plot for such multivariate data.

Figure(4.62): Example of typical quantile-quantile plot for distribution of *test*-values vs. beta distribution for multiple regression

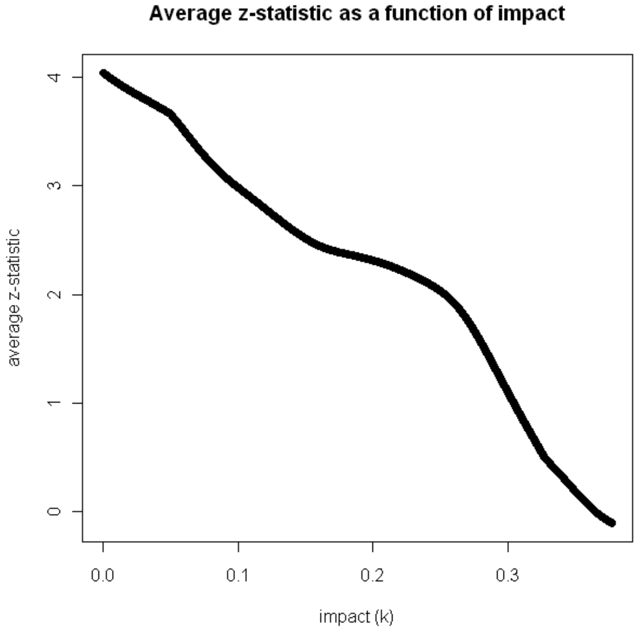


In summary, both the simple and multiple regression experiments indicated that the resulting distribution of *test*-statistics from controlling for an unobserved confounding variable is well approximated by a Beta distribution. In particular, as the first two moments of these distributions tend to be identical, the Beta distribution is a particularly relevant distribution from which to approximate and understand the dispersion of such inference statistics.

Such alignment of the distribution of estimated *test*-statistics with the Beta distribution prompts further suggestions. For instance, because the Beta distribution has a finite range, it is possible that the impact an unobserved variable with given Δ- and Γ-relationships has on an inference is limited. In other words we may be able to limit the maximum change in the test statistic for a given pair of given Δ- and Γ-relationships. This suggests a certain maximum impact of a confounding variable (MICV). For example, under such results an analyst could find the maximum impact threshold of a confounding variable (MITCV). Such an approach would suggest that although a variable with relationships equivalent to the MITCV would rarely invalidate an inference, it would represent the most conservative estimate of the robustness. Similarly, one could also speak of the minimal impact an unobserved confounding variable might have for a given Δ- and Γ-relationship. Together the average, minimum and maximum thresholds may present a more holistic understanding of how inferences might change when controlling for an unobserved variable.

Next, I examined the whether the average *test*-statistic continued to decrease as one increases the impact or product, *k,* of the Γ- and Δ-relationships as it did in the simple regression case. The results of the experiment with multiple regression indicated a very similar pattern in that as the impact increases the average *test*-statistic decreases. This result supplies the basis for an algorithm to identify the AITCV in that the algorithm can crawl along the curve until it reaches the desired *test*-statistic (e.g. 1.96). Figure (4.63) illustrates the estimated average *test*-statistic as a function of the impact for a given dataset.

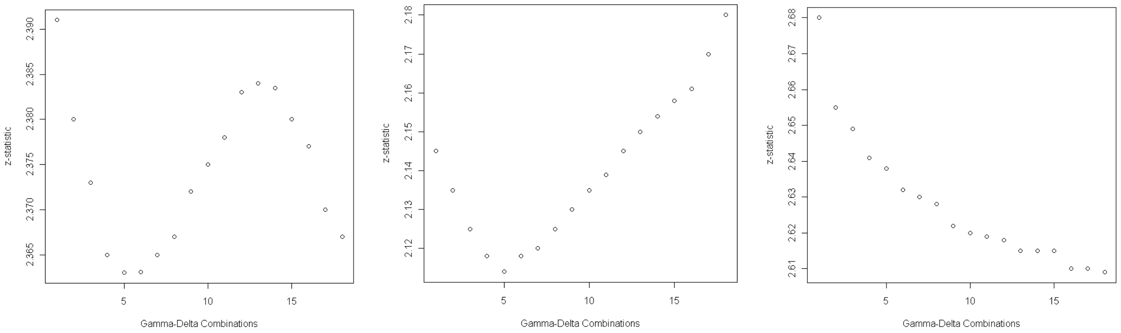
Figure(4.63): Average *test*-statistic as a function of the impact (product of Γ and Δ)



Finally, I assessed how the split of *k* influences the average *test*-statistic in multiple regression. In particular, I examined the influence of the split of *k* by considering impacts from 0.01 to 0.50 in increments of 0.01. Subsequently, for each fixed value of *k* I allowed the Γ-relationship to take on values from 0.05 to 0.90 in increments of 0.05 and assigned the Δ-relationship to quotient of *k*/Γ. I then estimated the average *test*-statistic for each Γ- and Δ- pair for a fixed *k* to understand how the split of *k* influences the average *test*-statistic. The influence of such splits tended to take on three general shapes. The first shape was similar to the cubic or sine curve whereas the second and third resembled more of a quadratic and asymptotically decreasing function (Figures (4.64) (a-c).

Figures (4.64) (a-c): Split of Impact Curves

(a) (b) (c)



However, regardless of which shape the split of the impact took on, the split of *k* had relatively little influence on both the range of *test*-statistics and the average *test*-statistic. In particular, how *k* was broken up into Γ and Δ tended to only affect the second decimal place of the average *test*-statistic. That is, although the split of *k* into Γ- and Δ-relationships does slightly widen the intervals and change the averages, it practically has little effect as it only affects the second decimal place of the average *test*-statistic. In other words, it is the product of the Δ- and Γ-relationships that drives the impact of a confounding variable on an inference statistic rather than solely one relationship. Such results support the need to consider both relationships simultaneously in assessing the sensitivity of a treatment effect.

Sensitivity of Results

To provide insight as to the sensitivity of the result in each of the four aspects discussed, I conducted sensitivity analyses. These analyses were carried out by holding all other parameters constant at their default values while a single parameter was varied. I broke these analyses into two strands, sensitivity to study characteristics and sensitivity to structural model choices. In particular, in examining the sensitivity to study characteristics I varied the probability of success from the default of 0.5 to 0.9 and 0.1 and varied the sample size from the default of 100 to 500 and 25. Further, I examined the sensitivity of the results to the assumed unconditional distribution of the unobserved confounder. In particular, I assessed how the results might change when assuming *U* comes from a Beta distribution. In terms of structural changes, I examined how the results varied when the true link function was a probit and we used the probit link. Further, I examined how results would vary when the true link function was the probit but we naively utilized the logit link.

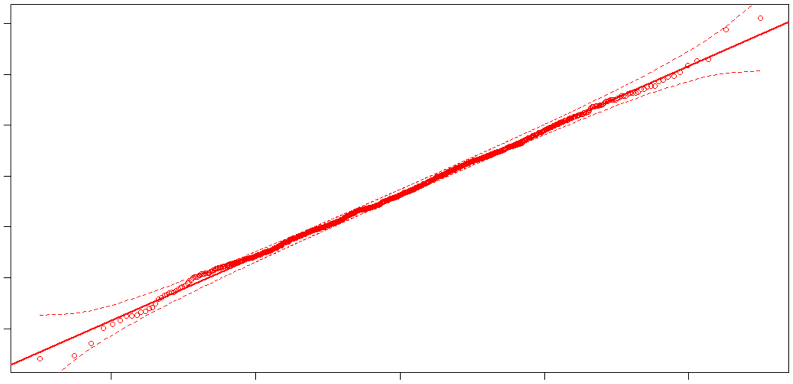
In the first set of these sensitivity analyses, I assessed on how previous results might change as a function of the probability of success, sample size and unconditional distribution of *U*. The results are summarized in Table (4.65). To a large extent, we saw little change as we increased or decreased the probability of success in the outcome. However, there were two noticeable changes. First, when the probability deviated from 0.5, the average distance of a standard deviation and thus the corresponding plausible *test*-statistic intervals tended to widen. In particular, through simulations, I estimated that on average the intervals would widen by approximately 10-15%. However in some specific instances the intervals widened up to 100%. In other words, although the coverage remained similar, the new distribution of *z*-statistics tended to me more dispersed than compared to when the probability of success was 0.5. Second, the approximation of the distribution of possible *test*-statistics by a Beta distribution was slightly degraded. In particular, though we rejected the null hypothesis that the distribution of *test*-statistics resulting from the impact of a confounding variable was near the nominal level of 5% when the probability of success was 0.5, with probabilities deviating from 0.5 we can now reject the null hypothesis between 7-11% of the time. In other words, when the probability markedly differs from 0.5 (e.g. 0.1 or 0.9) the approximation of the possible *test*-statistics by the Beta distribution is slightly less accurate. Next I examined how sample size may influence the interval sizes and approximate distribution. First, when decreasing the sample size to 25, we saw that the impact of a confounding variable on the treatment's *test*-statistic tended to be more dispersed. That is, the effect of an unmeasured confounding variable varies more widely as sample size decreases as the standard deviation was increased, on average, by 50%. The approximation of the resulting distribution of the *test*-statistics by the Beta distribution for higher sample sizes appears similar to that of original sample size as illustrated in Figure (4.66) Further, with increased sample size the *chi*-squared goodness of fit test was able to reject the distribution of the resulting *test*-statistics was a Beta distribution 5.5% of the time. However, in contrast, the *chi*-squared test suggested that we can reject the alignment of the distributions approximately 25% of the time when we reduce our sample size to 25. Next, when we randomly draw the values for *U* from a Beta distribution as opposed to a Uniform distribution, our results again remain similar to those prior. In particular, we only saw a slight increase (10%) in the magnitude of a standard deviation. Further, for each study characteristic change the coverage of the true *test*-statistic by the estimated distribution remained virtually unchanged. That is, approximately 68% and 96% of the time the true *test*-statistic fell within one and two standard deviation respectively.

Table(4.65): Examples of the average size of a standard deviation for the estimated *test*-statistic distribution of the treatment effect when adjusting for a unmeasured confounder in the presence of multiple measured confounders

Correlations			Original	Change in Prob		Change in Sample Size		*U*~Beta(α,β)
			Ave size	Ave size of SD		Ave size of SD		Ave size
Φ	Γ	Δ	of SD	Decrease	Increase	Increase	Decrease	of SD
0.3	0.1	0.1	0.021	0.016	0.021	0.015	0.019	0.020
		0.5	0.034	0.048	0.050	0.047	0.044	0.045
		0.7	0.038	0.050	0.049	0.045	0.079	0.049
	0.5	0.1	0.046	0.092	0.085	0.080	0.083	0.079

		0.5	0.046	0.063	0.059	0.082	0.093	0.055
		0.7	0.048	0.069	0.070	0.090	0.092	0.059
	0.7	0.1	0.090	0.080	0.089	0.093	0.099	0.080
		0.5	0.086	0.089	0.088	0.085	0.099	0.088
		0.7	0.099	0.102	0.100	0.098	0.105	0.101
0.5	0.1	0.1	0.010	0.016	0.015	0.018	0.021	0.019
		0.5	0.022	0.031	0.030	0.034	0.082	0.034
		0.7	0.057	0.075	0.080	0.067	0.071	0.069
	0.5	0.1	0.050	0.079	0.080	0.049	0.115	0.065
		0.5	0.080	0.091	0.090	0.079	0.108	0.089
		0.7	0.065	0.087	0.085	0.063	0.093	0.080
	0.7	0.1	0.065	0.069	0.067	0.069	0.089	0.067
		0.5	0.095	0.097	0.094	0.099	0.121	0.094
		0.7	0.056	0.072	0.073	0.055	0.111	0.069
0.7	0.1	0.1	0.013	0.015	0.010	0.019	0.019	0.015
		0.5	0.028	0.031	0.03	0.034	0.059	0.031
		0.7	0.063	0.065	0.063	0.069	0.080	0.064
	0.5	0.1	0.050	0.067	0.060	0.061	0.075	0.060
		0.5	0.045	0.059	0.063	0.044	0.057	0.056
		0.7	0.069	0.070	0.077	0.076	0.099	0.071
	0.7	0.1	0.118	0.130	0.101	0.111	0.123	0.112
		0.5	0.092	0.094	0.092	0.090	0.110	0.093
		0.7	0.067	0.070	0.069	0.069	0.089	0.070
Average			0.058	0.068	0.066	0.064	0.083	0.064

Figure(4.66): Quantile-quantile plot of Beta distribution versus resulting *test*-statistics when sample size is 25



In the second set of sensitivity analyses, I assessed the sensitivity of the framework to the choice of link function. In particular, I examined how the results change when the true link function is a probit rather than a logit and we use a probit link. Further I assessed how the results might change when the true link is a probit but we use a logit link. For the estimated distributions of the *test*-statistics, Table (4.67) displays the average size of the standard deviations and the estimated coverage for such distributions. In particular the first three columns specify the weighted partial correlations between the logit of the outcome and the treatment, Φ, between the treatment and the unobserved confounder, Γ, and between the logit of the outcome and the unobserved confounder, Δ. The next set of columns describes the original standard deviations and coverage for when the logit link was both used and was the proper link. The third set of columns describes similar standard deviations and coverage for when the probit link was both used and was the proper link. The final set of columns describes the same properties but when the true link function is the logit, however an analyst has used the logit link instead.

Table(4.67): Sensitivity of coverage of true test-statistic by estimated test-statistic distribution (proportion) and average standard deviation of estimated *test*-statistic distribution when there are multiple measured confounders

Correlations			Original (Logit for Logit)			Probit for Probit			Logit for Probit		
			Ave size of SD	Coverage in ± SDs		Ave size of SD	Coverage in ± SDs		Ave size of SD	Coverage in ± SDs	
Φ	Γ	Δ		1SD	2SD		1SD	2SD		1SD	2SD
0.3	0.1	0.1	0.021	0.677	0.953	0.016	0.677	0.955	0.019	0.677	0.956
		0.5	0.034	0.664	0.952	0.042	0.690	0.952	0.048	0.680	0.956
		0.7	0.038	0.687	0.952	0.067	0.680	0.958	0.071	0.679	0.959
	0.5	0.1	0.046	0.693	0.950	0.090	0.677	0.955	0.086	0.671	0.957
		0.5	0.046	0.678	0.958	0.086	0.684	0.952	0.089	0.680	0.955
		0.7	0.048	0.676	0.954	0.097	0.685	0.953	0.068	0.674	0.957
	0.7	0.1	0.090	0.674	0.950	0.099	0.683	0.953	0.143	0.693	0.946
		0.5	0.086	0.666	0.960	0.113	0.687	0.955	0.102	0.680	0.959
		0.7	0.099	0.685	0.948	0.081	0.689	0.954	0.097	0.688	0.955
0.5	0.1	0.1	0.010	0.690	0.946	0.015	0.681	0.955	0.020	0.676	0.957
		0.5	0.022	0.701	0.950	0.041	0.695	0.952	0.051	0.689	0.955
		0.7	0.057	0.684	0.957	0.068	0.682	0.959	0.075	0.682	0.955
	0.5	0.1	0.050	0.657	0.963	0.091	0.670	0.955	0.080	0.669	0.951
		0.5	0.080	0.677	0.953	0.085	0.672	0.953	0.088	0.671	0.954
		0.7	0.065	0.665	0.959	0.082	0.681	0.952	0.072	0.673	0.950
	0.7	0.1	0.065	0.689	0.956	0.099	0.672	0.953	0.132	0.690	0.951
		0.5	0.095	0.658	0.960	0.114	0.689	0.956	0.100	0.681	0.956
		0.7	0.056	0.675	0.955	0.080	0.690	0.954	0.092	0.682	0.949
0.7	0.1	0.1	0.013	0.671	0.952	0.015	0.675	0.957	0.019	0.673	0.957
		0.5	0.028	0.696	0.950	0.044	0.688	0.954	0.054	0.687	0.951
		0.7	0.063	0.658	0.959	0.067	0.685	0.956	0.065	0.669	0.958
	0.5	0.1	0.050	0.657	0.963	0.090	0.673	0.954	0.061	0.674	0.957
		0.5	0.045	0.683	0.953	0.087	0.679	0.955	0.089	0.680	0.947
		0.7	0.069	0.669	0.956	0.092	0.684	0.956	0.087	0.669	0.957
	0.7	0.1	0.118	0.661	0.964	0.098	0.689	0.955	0.098	0.682	0.956
		0.5	0.092	0.662	0.962	0.111	0.691	0.955	0.140	0.679	0.959
		0.7	0.067	0.691	0.958	0.084	0.689	0.954	0.078	0.690	0.954

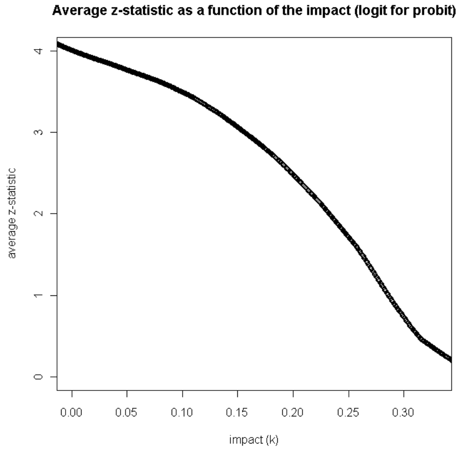
Average	0.058	0.676	0.955	0.076	0.683	0.955	0.079	0.679	0.955

The results suggested that the AITCV framework, as applied to the models with the logit link, has similar success when applying the framework to models using the probit link. In particular, though the average size of the standard deviation of the distribution of the potential *test*-statistics grows from an average of 0.058 to 0.076, the coverage remains almost identical. Further, when misapplying the logit link rather that the true probit link, we again found the size of the standard deviations to be slightly inflated but coverage remain the same.

I found similar evidence in examining the approximation of the estimated *test*-statistic distribution by the Beta distribution. In particular, when we properly used the probit link the *chi*-squared test for the null hypothesis that the *test*-statistic distribution was a Beta distribution, I was only able to reject the null hypotheses 5.6% of the time. In comparing this with using the logit link properly (5.7%), both are only marginally above the nominal α-level of 0.05 and are nearly identical. In contrast, when improperly utilizing the logit link for the proper probit link, my simulations suggested that the *chi*-squared test rejected the null hypothesis 8.6% of the time. In other words, the distributions of *test*-statistics resulting from the impact of a confounding variable with said Γ- and Δ-relationships is well approximated by the Beta distribution for both the logit and probit links and is slightly weakened when we have incorrectly specified the link function.

Similar to prior simulations, I examined the whether the average *test*-statistic continued to decrease as one increases the impact or product, *k,* of the Γ- and Δ-relationships as it did in the simple and multiple regression case. The results of the experiment with both changes in the link function and changes in the study characteristics indicated that as the impact increases the average *test*-statistic decreases. Figure (4.68) displays a pattern very similar to that of the logit case. Last, I reassessed how the split of *k* influences the average *test*-statistic when we make such changes to the simulations. Again, the split of *k* had relatively little influence on both the range of *test*-statistics and the average *test*-statistic indicating it is the product of the Δ- and Γ-relationships that drives the impact of a confounding variable on an inference statistic.

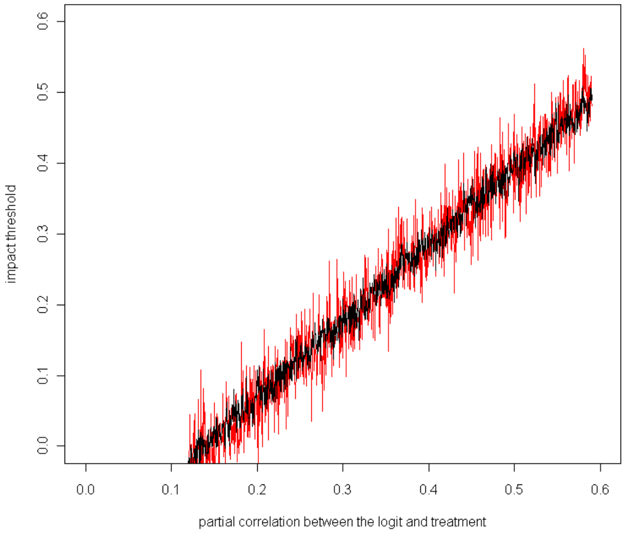
Figure(4.68): Average *test*-statistic as a function of the impact when the logit link is misapplied in place of the correct probit link



AITCV

Finally, I applied the AITCV framework to estimate the average impact threshold of a confounding variable at the nominal α-level of 0.05. In particular, I utilized the AITCV framework in conjunction with the monotonic relationship between the average *test*-statistic and the impact, *k*, to identify thresholds corresponding with p-values of 0.05 for given datasets. The results indicated that on average as the partial correlation between logit of the response and the treatment increases the impact needed to invalidate that inference increases. However, the AITCV in binomial regression models is somewhat dependent on the characteristics of each particular dataset. In other words, datasets with identical correlation matrices may have slightly different AITCV depending on the each study's characteristics. Despite some dependency on study characteristics, the fluctuation in AITCV tends to decrease as the number of measured control variables increases. That is, the more variables we measure and include in our analysis, the more stable our AITCV becomes. For instance, Figure (4.69) displays the AITCV as a function of the partial correlation between the logit of the response and the treatment.

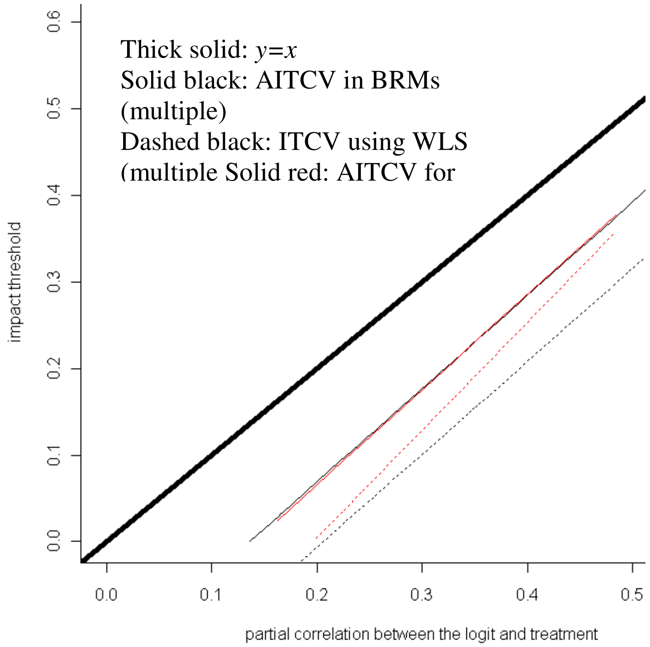
Figure(4.69): Variation in AITCV for datasets that use no control variables (red) and nine control variables (black)



In particular, plotted in red are the AITCV's for various datasets which do not use any control variables. In contrast, plotted in black are the AITCV's for various datasets which use nine control variables in the BRM. Evident from the figure is the increased robustness of inferences stemming from those datasets with higher partial correlations between the logit of the response and the treatment. Further, the figure also illustrates the reduced dispersion of AITCV when controlling for multiple covariates. In other words, retaining relevant control variables in estimating a treatment effect not only adjusts your treatment effect estimate for imbalances and makes it more robust to unmeasured confounding variables but also reduces the range of the AITCV. Such a property demonstrates a type of secondary absorption in that using control variables not only potentially absorbs the impact of a confounding variable but also partially absorbs the range of impacts a confounding variable can have. As a result, including measured covariates related to the response helps a researcher better identify both the treatment effect and the AITCV.

In addition to understanding how the AITCV changes as both a function of the response-treatment partial correlation and the number of control variables used, it can also be useful to compare the AITCV in BRMs with the ITCV in WLS models. Specifically, it can be informative to understand the difference between the thresholds. Figure (4.70) displays the four separate lines representing the average AITCV's for a given response-treatment partial correlation in addition to the thick black *y*=*x* references line. Starting from the bottom we have a dashed black line which represents the ITCV using WLS with no control variables. Next, we have a dashed black line representing the ITCV using WLS multiple control variables. Next we have a solid red line representing the AITCV for no control variables. Finally we have the solid black line representing the AITCV when controlling for multiple variables.

Figure(4.70): AITCV in BRMs (solid) versus the ITCV in WLS (dashed) for no (red) and multiple control covariates (black)



From this figure we can see that impact thresholds of BRMs tend to be higher than the corresponding thresholds in WLS. That is, for a given response-treatment correlation, inferences drawn from BRMs tend to be more robust than there WLS counterparts. I suggest two reasons for such increased thresholds. A first reason such thresholds are lower is that they are additionally influenced by the estimates of the dispersion parameter. As BRMs constrain the dispersion parameters to be one, we must adjust the standard errors by a factor equal to the estimated dispersion (e.g. (4.29)). Such adjustments affect the corresponding *test*-value and thus the thresholds. A second reason why such thresholds are higher on average, is the iterative nature of estimating the maximum likelihood estimate in BRMs. If we recast the Fisher scoring method as a IRWLS problem as above, it becomes evident that the adjusted dependent variable is pulled toward a linear combination of the variables in the model at each iteration. In other words if one has ten variables in a model and then considers an eleventh variable, the influence of that variable is constrained to some extent by its strength compared to the other ten in predicting the response. In contrast the maximum likelihood estimator in the linear model

(OLS or WLS) uses a single iteration and there is no need to form an adjusted dependent variable. As a result, the dependent variable is stable which allows for greater influence of a single variable. In other words, maximum likelihood estimates in BRMs are more dependent on the variables in the model because they require numerical estimation.

Example for Multivariate Extension

To further illustrate how the AITCV may be applied in a given context, I extend the simple BRM example provided above to include a control variable. Whereas the first example estimated the unconditional association of father's education with attaining the minimum literacy requirements for advancing to the next grade, this extension focuses on conditional comparisons. The focus of this illustrative example remains on the relationship between reading achievement and father's education level for sixth grade students in South Africa's Limpopo region. However, I now posit that mother's education level is also associated with the child's attainment potentially represents a measured confounding variable. More specifically, I first examined the BRM

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1_m}(fathers) + \beta_2(mothers) \qquad (4.71)$$

where *fathers* and *mothers* represents the educational level of the respective parent. Though father's education level is likely associated with his child's educational attainment, it is likely that in adjusting for mother's education level the estimate of father's education will diminish. Using the same SACMEQ data for the Limpopo region of South Africa, Table (4.72) presents the estimates of father's education while holding mother's education level fixed.

Table(4.72): Regression of Minimum Educational Competency on Father's and Mother's Education Level

	Coefficient	Standard Error	Z-value
Intercept (β_0)	-2.90	0.37	-7.85
Father's Education (β_1)	0.19	0.09	2.19
Mother's Education (β_2)	0.23	0.08	2.73

Evident from contrasting Tables (4.40) and (4.72), the estimate of the relationship between father's education level and minimum literacy competency decreases as does the test statistic though it remains significant. In other words, though it is unclear how mother's education might confound father's education, there is at least some overlap or collinearity between mother's and father's education. Now in conducting a sensitivity analyses we ask what must be the magnitudes of the Δ- and Γ-relationships an unmeasured confounding variable must possess in order to alter the significant inference for father's education while controlling for mother's education. That is, the question of interest focuses on whether the statistical inference based on (4.71) would be altered when controlling for other possibly unknown factors.

To assess the robustness of the original inference using the AITCV framework I first estimated the weighted correlations between the logit of the binomial proportions, father's education and mother's education (Table (4.73)) using the weight from the final iteration of (4.71).

Table(4.73): Weighted Correlations of Variables in (4.71)

	Father's Education	Mother's Education	Logit of Binomial Proportion
Father's Education	1		
Mother's Education	0.45	1	
Logit of Binomial Proportion	0.56	0.56	1

Next, I arbitrarily assigned a starting value of 0.5 to both the Γ- and Δ-relationships where they now represent the partial correlations with the unknown confounding variable according to (4.44), (4.45) and (4.46). Next I randomly drew a number from the Uniform distribution to represent the correlation between the control variable, mother's education and the unknown confounding variable whose bounds are estimated by (4.52) and are -0.73 and 0.91 in this example. Using (4.50) and (4.51) I estimated the zero-order correlation matrix. Table (4.74) presents the zero-order correlation matrix between the variables in this analysis where Γ^{zero} and Δ^{zero} represent the zero-order correlation based on Γ and Δ.

Table(4.74): Weighted Correlations of Variables in (4.71) Plus a Confounding Variable

	Father's Education	Mother's Education	Logit of Binomial Proportion	Unmeasured Confounder (*U*)
Father's Education	1			
Mother's Education	0.45	1		
Logit of Binomial Proportion	0.56	0.56	1	
Unmeasured Confounder (*U*)	Γ^{zero}	U($\rho_{umother}^{(lower)}$, $\rho_{umother}^{(upper)}$)	Δ^{zero}	1

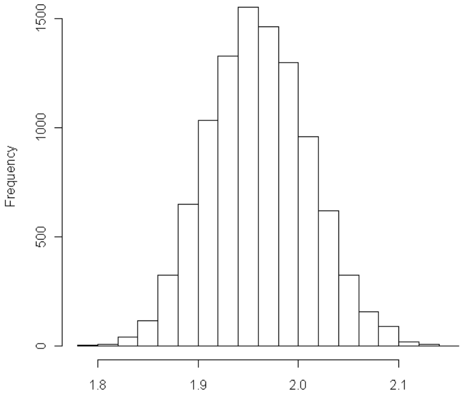
Based on this correlation matrix, I generated a single unmeasured confounding variable, *U*, and re-estimated the effect and significance of father's education on educational attainment using

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{1_m}(fathers) + \beta_2(mothers) + \beta_3(U^{(1)}) \qquad (4.75)$$

Repeating this process 1000 times, including a new random draw from U($\rho_{umother}^{(lower)}$, $\rho_{umother}^{(upper)}$), I estimated how the *test*-statistic of father's education would change if we had controlled for a confounding variable with an impact of 0.25 or Γ- and Δ-relationships equal to 0.5. Taking the average of the 1000 *test*-statistics, I estimated that if we had controlled for a confounding variable with such relationships the average *test*-

statistic would be 0.93. Based on prior simulation results concerning the monotonic relationship between the average *test*-statistic and the impact, this suggests that the AITCV is lower than 0.25. As a result, I next adjusted the impact and corresponding thresholds to something less than 0.25. I repeated this process until the average *test*-statistic produced was at the critical value of 1.96. Such simulations suggested that the AITCV for the data at hand was 0.05 or roughly 0.22 for both the Γ- and Δ-relationships. That is, in order to alter our original inference we would need a confounding variable with Δ- and Γ-relationships roughly equal to 0.22. Though a variable with such Γ- and Δ-relationships on average decreases the test-statistic of father's education to 1.96, the effect of a confounding variable with such relationships takes on a range of values as discussed earlier and illustrated in (4.76).

Figure(4.76): Histogram of Simulated *test*-statistics for Regression of Minimum Educational Competency on Father's Education, Mother's Education and a Hypothetical Confounding Variable



As demonstrated in the figure, a confounding variable with Γ- and Δ-relationships of 0.22 reduces the test statistic of father's education from 2.19 to 1.96 on average. However, less frequently a confounding variable with those relationships reduces the *test*-statistic to as low as 1.8 or as high as 2.1. In other words, according to the simulation 68% of the time the *test*-statistic for father's education level when controlling for mother's education and the hypothetical confounding variable, *U*, changes to a value between 1.91 to 2.01. Similarly, 96% of the time the test-statistic changes to a value of 1.85 to 2.06 and 99% of the time to a value of 1.80 to 2.11. Similarly, if we decided to be highly conservative, we might ask what the MITCV is. In other words what is the minimum Δ- and Γ-relationships needed to invalidate the inference? Again, such a question represents the most conservative approach as it accepts that such Δ- and Γ-relationships would invalidate an inference only in the most extreme circumstances. In this particular data, the MITCV would be 0.03 or roughly 0.17 for the Δ- and Γ-correlations. Such values taken together with the AITCV present a more holistic picture. In other words, in assessing the robustness of this inference, we can now say that on average an unobserved confounding variable would need to have correlations with the treatment and outcome of approximately 0.22. However, in some rare circumstances an unobserved confounding variable with correlations as low as 0.17 with the outcome and treatment may invalidate our inference. Again, using either the AITCV or the MITCV I do not speak to the likelihood of such a confounding variable existing, however such an approach does quantify the magnitude of relationships needed by confounding variable to indicate that father's education is not significantly associated with educational attainment when controlling for mother's education.

Notes to Chapter IV

[1] One could additionally consider a linear probability model. In such cases assessing the robustness of inferences could be directly accomplished by the original Impact Threshold of a Confounding Variable framework developed in Frank (2000).

[2] Also of interest may be the robustness of an inference to multiple unobserved confounding variables. In such a case one might suggest the true model as

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 Z + \beta_2 X_1 + \beta_3 X_2 + ... + \beta_{p+1} X_p + \beta_{p+2} U_1 + ... + \beta_{p+2+j} U_j \qquad (4.77)$$

Though such extensions are simple in the OLS linear model framework as they reduce to partial correlations, in the BRM framework such extensions may not be trivial. In particular, because in BRMs we solve for maximum likelihood estimates using iterative optimization methods, the adjusted dependent variable is constructed using a linear combination of the predicted variables. As a result, one can not necessarily reduce multiple unobserved confounding variables into a single unobserved confounding variable and obtain the same estimates or influences. This aspect is suggested for future research in the final chapter. However, if one were to use a linear probability model, the methods presented in Frank (2000) would be directly applicable. Alternatively, one could also view all unobserved confounding variables as measurements of a single unobserved latent confounding variable.

CHAPTER V

CONCLUSION

I have focused this dissertation on improving causal inferences in common educational settings. Such developments focused on two pivotal covariate relationships; the covariate's unique relationship with the outcome and its unique relationship with the treatment. In my first study I argued that in educational settings one must frequently take into account the influence of the group in understanding both treatment assignment mechanisms and effects of such treatments. In other words, because teaching and learning are inherently part of a multilevel system, almost every educational intervention or treatment is likely influenced or mediated by group membership. Further, because of the complexity of these systems, I argued that it is likely that such influence will be heterogeneous among groups and among individuals within the group. As a result, in order to assume the strong ignorability of the treatment assignment in educational settings, we may frequently need to construct groups that are balanced on pretreatment covariates at all levels of the teaching and learning process. That is, assessing the effects of treatments without taking into account such influence may provide biased inferences and incomplete understandings of the teaching and learning processes.

In accepting the potential for groups to exert influence over individuals' treatment assignments, one is further challenged in specifying the manner in which the group effectively influences the individual's treatment assignment. The challenge of conceptualizing such treatment assignment mechanisms in multilevel settings arises because different groups may have very different influences over their respective members. Further, even within such groups, a group's role in predicting the treatment assignment may vary among subgroups of individuals. To address the differential roles groups take on, I considered several manners by which groups may influences their respective individuals and suggested that appropriate identification of these mechanisms is important to effective and efficient estimation. My results suggested that while consideration of the differential roles groups take on in predicting treatment assignment was important, the model one uses to estimate such influence matters to a lesser extent. In other words, I claimed, in this first study as well as the others, that it is more crucial to include the appropriate variables in a model than to use a specific model. Such claims, however, were paired with additional complications. In particular, identifying 'appropriate' variables to include in a PS model required some criterion by which to judge the inclusion of a variable.

To advance understanding of which variables might be included in a PS, my second study developed an approach to select the most effective and efficient variables in observational studies resembling cluster randomized trials. I argued that despite the bias reduction one receives from approaches such as including all available variables, the unidimensional focus on bias frequently degrades the quality of the estimator. For instance, returning to an earlier example, estimating a treatment effect with a single person represents an unbiased estimate of the true treatment effect. However, such an approach is unrealistic as it tends to be inefficient and highly variability. For this reason, I focused this study on developing a strategy that would balance the bias and the variance of the treatment effect estimator. Further, to extend support for the dominance of variable measurement and selection over model type, both the first and second studies assessed the efficacy of different PS uses. Specifically, in both studies I assessed the performance of using the PS as a stratification criterion, a case weight criterion and as a matching criterion. The results showed little difference among the three uses and thus lend credence, again, to the importance of measuring and adjusting for the appropriate variables over type of PS model or PS use.

Maintaining focus on the importance of measuring and adjusting for the appropriate variables, in the final study I developed a method to assess the robustness of one's inferences to an unmeasured confounding variable. In particular, the method attempted to quantify the magnitude of the two pivotal relationships the unmeasured confounding variable would need to have to invalidate the original inference in the context of binomial regression models (BRMs). In this study, I again assessed the role of model type within this framework. Consistent with the results of the first two studies, the study suggested that model type had very minimal influence over the results. For instance, using a probit link function as oppose to logit link function had next to no influence on the impact thresholds. The consistency of results among the three studies despite somewhat divergent settings supports the relative importance of the measurement and inclusion of appropriate variables as compared to the model type or PS use. Further, such consistency may be evidence to suggest the importance of measuring the appropriate variables in a variety of other situations.

Substantively, the papers explored a range of important factors in education including teacher literacy knowledge, retention policies, and parental education. Such factors emphasize the multilevel nature of the schooling in that they consider multiple sources of inputs. That is, the studies presented in this dissertation collectively explored the effect of individual student influence (i.e. parental education), shared teacher/classroom influence (i.e. teacher knowledge) and shared school influence (i.e. retention policies). The first study which examined the effect of teacher literacy knowledge on literacy achievement suggested a small but persistent and practically significant effect on reading comprehension. In other words, it is likely that one avenue by which teachers influence student achievement is teacher knowledge. However, because teacher knowledge and quality is likely influenced by or concentrated in certain schools, understanding its complex role likely requires deconstructing the school's part in attracting, developing and retaining teachers with high knowledge. As a result, such measures of teacher quality will likely be a high priority as legislation continues to require highly qualified teachers for every student. The second study, which examined the effect of school retention policies on the average student achievement, offered findings concerning the influence of a school level factor. In particular, the second study suggested that there was no noticeable different in overall student achievement when comparing schools that allow retention and those that do not. Finally, the third study suggested that an important student level factor that contributes to student achievement is parental education. Though each study took on very different constructs at different levels of the schooling system, each study demonstrated the potential of factors external to the student in influencing student achievement. Such factors have been, and will likely remain important in understanding the processes that contribute to student learning.

In applying the methodological results to similar or alternative nonrandomized designs, questions arise about the generalizability of these results and implications. That is, how might the individual results (e.g. include variables that exceed the Δ to Γ/2 ratio in the PS) as well as the collective results (e.g. the inclusion of appropriate variables in the model dominates the choice of model type) generalize to other situations? Application of these methods to other designs and studies may have several sources of variation to consider. Perhaps a first source of such variation is the study design elements. Specifically, these studies assessed various aspects of estimation and inference within the context of specific designs common to educational data and used a limited set of parameters. For instance, in the study concerning multilevel propensity scores, extending the results to other studies may at a minimum produce variation in the findings as a product of the study's exclusive focus on multi-site randomized trial designs with only six variables. Similarly, the other studies concerning PS construction and robustness of inferences focused on designs resembling cluster randomized trials and individually randomized trials, respectively. Each of these designs retained idiosyncrasies that may not be present in other designs. A second potential source of variation in these studies may be the treatment type. The first and last studies expanded the type of treatment to include both dichotomous and continuous treatments whereas the second study exclusively examined dichotomous treatments. However, each of these studies exclusively considered treatments with homogeneous effects. Relaxing such constrictions may produce additional variations in the results.

A third possible source of variation in these studies is the choice of analytic methods. In the variable selection study, I exclusively focused on a logistic model to estimate the PS’s and a hierarchical linear model (HLM) with a random intercept to estimate the outcome model. It is possible that the estimated thresholds vary depending on choice of PS and outcome model. For instance, thresholds plausibly change when one considers a fixed effect model or uses nonparametric methods. Analogous, the remaining studies also imposed analytic constraints. The multilevel propensity score study strictly utilized hierarchical (generalized) linear models with an identity or logit link to estimate the propensity score. Similarly, the last study focused mainly on the BRMs with a logit link and to a lesser extent investigated the probit link. Varying such analytic constraints may also produce variation in the results.

Despite the potential variation of such results in alternative designs and situations, the individual and collective results of these studies suggest a number of implications. Returning to the purposes of this dissertation, I partition the implications of each study along three dimensions of causal inference: study design, analysis and inference.

A number of implications for the prospective and retrospective design of nonrandomized studies in education can be drawn from largely the first two papers but also from the last paper. Here, I use prospective design to refer to the planning and design phases of a study that take place before any data is collected. In contrast, I refer to the retrospective design phases as those that take place after the data has been collected which attempt to approximate a randomized experiment. Within the context of the first study which addressed multilevel treatment assignment mechanisms, the study suggested that researchers carefully attend to the role of group memberships and characteristics in constructing comparable groups of individuals. That is, if the treatment assignment is likely influenced by group membership or clustered among individuals within the same group, ignoring the influence of the group may bias estimates. As a result, an initial implication for the prospective design of observational studies is that researchers should emphasize accurately measuring relevant variables at all levels that potentially influence the treatment assignment. However, the first two studies further developed and modified such design implications. In particular, the studies demonstrated that although a variable at any level may directly influence the treatment assignment, its relevance in removing bias from the treatment effect estimator is constrained by its relationship with the outcome. In the framework of the Rubin Causal Model (RCM), one might link such constraints to ignorability of the treatment assignment or the conditional independence of the potential outcomes and treatment assignment. For instance, assume the treatment assignment mechanism only depends on two sets of variables, $\vec{X}$ and $\vec{W}$, where each variable in $\vec{X}$ is additionally related to the outcome (e.g. $\Delta \neq 0$) but each variable in $\vec{W}$ is (conditionally) unrelated to the outcome (e.g. $\Delta = 0$). Because $\vec{W}$'s conditional outcome-covariate relationships ($\Delta(\vec{W})$) are zero, the potential outcomes must be must be conditionally independent of the treatment assignment given $\vec{X}$. That is,

$$\Delta(\vec{W}) \approx 0 \Rightarrow Y^{(i)} \perp Z \mid \vec{X} \tag{5.1}$$

In other words, the treatment assignment is ignorable when controlling only for $\vec{X}$. In this way we can use the duality of confounding to our advantage within the RCM.

As a result of my analyses, a particularly relevant prospective and retrospective design implication emerges. When studying a treatment, we should first focus our efforts on collecting information on those variables related to the outcome. A further implication relevant for educational research is that one can still conduct high quality studies when studying relatively new treatments or treatments that not well understood as long as the outcomes in the study are well understood. For example, though teacher knowledge is an attribute that has been of interest for some time, teacher knowledge and the factors which influence it are not very well understood. This dissertation suggested that the quality of our estimates relies less on identifying and measuring those factors which predict teacher knowledge and more on identifying those factors that influence student achievement. In other words as student achievement is an outcome that has been well studied, the findings suggested that high quality estimates of the treatment effect are still obtainable despite the lack of understanding concerning teacher knowledge.

In the prospective design of a study, we might thus prioritize three classes of variables. First, we would like to measure those variables that have a strong relationship with the outcome regardless of their relationship to the treatment. Their measurement might be considered essential as they will at a minimum increase efficiency and likely decrease bias considerably. Second, we might focus our resources on those variables thought to be weakly related to the outcome. The utility of their measurement depends to a large extent on their relationship with the treatment. However, their measurement protects the study from unforeseen imbalances between the treatment groups which are especially important in treatments that are not well studied. The lowest priority variables are those that are historically unrelated to the outcome. Their value jointly depends on the magnitude of chance relationships with the outcome and their relationship with the treatment. In summary, rather than attend to the idiosyncrasies of a specific study's treatment selection process, prospective design should largely privilege measurement of variables that retain stable relationships with the outcomes.

In a similar manner, the first two studies suggest that retrospective design should focus on constructing meaningfully comparable groups rather than simply comparable groups. That is, in the retrospective design of a study, it is advantageous to design comparable groups such that it approximates the most effective and efficient randomized study one can with the data. This may frequently imply that one should try to mimic a block randomized study by blocking on characteristics with which the outcome covaries. Further, given the extended range of variables to consider in multilevel settings, the results also suggested a complementary implication for the retrospective design of a study. That implication advocates precaution concerning which variables to include in the PS model. That is, the realized benefits of measuring such hierarchical memberships and variables as outlined in the prospective design are dependent on the variable's respective empirical relationships with the outcome and treatment assignment. In particular, the second study implied that retrospective designs using the PS should construct the PS with the most effective and efficient covariates to improve the quality of their estimates in terms of MSE. A rough rule of thumb for the retrospective design of a study that emerged was to construct the PS using only those variables whose unique relationship with the outcome (Δ) is at least half of its unique relationship with the treatment (Γ). In other words, for those variables whose Δ-relationship is less than half of the corresponding Γ-relationship, the variable's contribution to the variance of the treatment effect estimator likely exceeds the amount of bias reduction it supplies.

Such results additionally underscore the importance of prospective designs that include outcome proxies such as a pretest measure. Prospective designs such as the pre-post test design allow one to retrospectively design and efficient and effective quasi-experiment by estimating the Δ-relationships without actually using the observed outcome. This salient feature of educational data combined with the efficient bias reduction value of the pretest stresses the importance of obtaining a pretest measure. That is, though literature has widely demonstrate the benefit of having a pretest measure, this study further adds to and complements such literature.

With respect to the analytic dimension of causal inference, the dissertation suggested several implications linked to those in the design dimension. First, the studies demonstrated that the similarities among various models tended to be greater than the differences among them. In other words, of the model types proposed in both the multilevel PS study and the AITCV study, each method had similar effectiveness in both removing bias and doing so in an efficient manner. In a similar manner, the studies also suggested strong similarities between PS uses. That is, weighting by the inverse probability of receiving the treatment, subclassifying or matching on the PS all tended to perform similarly. In contrast collectively the studies suggested the variables one includes in the PS may cause estimates to diverge considerably. In terms of implications for the analysis, these results suggest that quality of the estimator hinges more on the variables than the models. In other words, one should comparatively devote more time and resources to selecting variables than to selecting among several reasonable models. Accordingly, this emphasizes the role of the partnership between the substantive researcher and the statistician.

Implications for the final generic dimension of causal inference, inferences, may also be drawn from each study. In particular, similar to the previous studies and literature, the collective results suggested that the appropriate adjustment for measured confounding variables is crucial to the sensitivity of inferences for at least three reasons. First, their inclusion improves the quality of the estimator and associated inferences. Second, the inclusion of measured confounding variables can absorb the impact of unmeasured confounding variables (Frank et al., 2008). More specifically, the amount of influence an unmeasured confounding variable may have on an estimated treatment effect and its associated inference can be reduced or absorbed because of nonzero relationships of measured confounding variables with unmeasured confounding variables. Third, the adjustment for measured confounding variables additionally absorbs the impact of an unmeasured confounding variable by reducing the range of possible test statistics. That is, we saw absorption take on an expanded role in iteratively optimized models such as maximum likelihood estimates in binomial regression models. In other words, when one has not controlled for any confounding variables, the range of possible test statistics one would see when considering the impact of an unmeasured confounding variable tends to be larger than the range one would see when we had originally controlled for multiple measured confounding variables. The importance of such adjustments for multiple confounding variables illustrated in the last study has an implication for drawing inferences. That implication is that not only will one improve the quality of one's estimates and inferences by adjusting for measured confounding variables, but one will also increase the robustness of such inferences to the constant threat of unmeasured confounding variables.

Though these studies addressed several general research questions, their designs, analyses and inferential conclusions each focused on specific contexts. As a result, this research leaves room to explore in at least three directions. The first direction focuses on understanding the generality of the results within the specified contexts whereas a second direction focuses on understanding the applicability of the current results to other contexts. Further, a third direction remains to extend such results to more complex settings that view education more holistically.

The multilevel propensity score study attempted expand the scope of those factors which may influence the treatment assignment by assessing the different roles groups may have in influencing their members' treatment assignments. Despite the increased scope, such a view still may not attend to the complexities of the schooling process. For instance, as schooling is a longitudinal process, students tend to belong to different groups over time. Such memberships may additionally inform students or teachers treatment choices. Similarly, learning and teaching tend to extend beyond the borders of a school and are plausibly influenced by external factors such as those represented by family and neighborhood characteristics. As a result, a more comprehensive picture may conceptualize the probability of being assigned to or selecting a certain treatment assignment as a function of neighborhoods, family, school, teacher and district characteristics. To address this more holistic understanding, I suggest using partial random effects or cross classified models. However, the conceptual basis for such an approach as well as the effectiveness of such models has not been studied. Future research in this area may focus on developing and evaluating the feasibility of such approaches.

A further line of inquiry may examine the robustness of using fixed effects in estimating multilevel PS's. In particular, the fixed effects model specification in study one tended to outperform the random effects models when the true confounding variables were excluded. As a result, it illustrated a certain robustness to the omission of true confounding variables. Economic literature has long suggested this advantage of fixed effect models in traditional outcome models. However, a similar advantage in PS has not been studied to my knowledge. Future research assessing the generality of this protection in more complex situations may have widespread implications for PS models that consider group influence.

A line of research emerging from the second study may be the applicability of the rough Δ to Γ rule (i.e. include variables whose Δ-relationship exceeds half of its Γ-relationship) to other study designs and models. Moreover, in the context of this dissertation, the applicability of that rule in the context of treatments that consider group influence is particularly relevant. Understanding how PS covariate inclusion thresholds may differ among different classes of models and within such classes is particularly relevant in estimating treatment effects using PS based methods.

In addition, another relevant issue that arose in this study was the tradeoff between the bias and variance belonging to the proxy versus cross validation methods of estimating the outcome covariate relationships. In particular, future research might focus on the thresholds at which it becomes more effective to use cross validation than a proxy approach. Such research might focus on two different potential thresholds. The first is sample size. With an infinite sample size, theoretically the cross validation approach should dominate any proxy based approach. However, empirical research is done with finite sample sizes. Accordingly, identifying the sample size at which it becomes more effective to use a proxy based approach may be relevant for large and small sample studies. A second threshold of interest in educational data is the magnitude of the proxy's relationship with the outcome. In particular, in the current study if the proxy is correlated with the outcome at 0.70, the bias of the estimator tended to be overshadowed by the variance of the estimator. That is, a proxy correlated with an outcome at 0.70 tended to insert relatively little bias in to the estimates. However, we might consider other proxy or pretest measures that tend to be less correlated with the outcome. As a result, a relevant future issue in education is the point at which the relationship between the proxy and outcome become so low that the bias begins to dominate the variance of an estimator.

Finally, the impact threshold of a confounding variable framework has some natural extensions beyond BRMs. In particular, future research might extend the framework to the larger class of generalized linear models (GLMs). That is, one might consider outcomes with distributions in the exponential family. In addition, within the context of the exponential family of distributions, it is of interest to consider the case of multiple confounding variables. In other words, rather than assess the robustness of an inference to a single unobserved confounding variable, one might ask how thresholds change if there are multiple unobserved confounding variables. In the OLS linear framework, such extensions are trivial as one can reduce the multiple unobserved variables into a single variable and still preserve the relevant relationships and statistical properties (Frank, 2000). However, iteratively optimized estimates may not preserve such convenience as a result of the changing adjusted dependent variable. For example, in GLMs the adjusted dependent variable is progressively pulled toward a linear combination of the variables. Such drift of the adjusted dependent variable toward other variables may alter those convenient properties seen in the OLS model. A final area for future research in this line is to understand the maximum and minimum impact an unobserved confounding variable can have on an inference for a given Δ- and Γ-relationship. That is, because the final study suggested that the distribution of possible test statistics is well approximated by a Beta distribution, future research may try to show that for a given Δ- and Γ-relationship the change in the test statistic is bounded. In other words, because the Beta distribution has finite bounds, future research might explore whether those bounds also apply to how the test statistics changes when controlling for an unobserved confounding variable.

APPENDICES

APPENDIX A

Derivation of Multilevel Propensity Score for Continuous Treatments

Following the logic and notation of (Hirano & Imbens, 2004), assume *Z* is continuously distributed measure, $\{Y_i^{(z)}\}\, for\ z\, in\, Z$, *Z* and ***X*** are defined on a common probability space and that $Y = Y(Z)$ is a well defined random variable (Hirano & Imbens, 2004). Further assume the ignorability of the treatment assignment such that

$$Y_i^{(z)} \perp Z \mid \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r} \tag{5.2}$$

Define unit *i*'s outcome *Y* and treatment assignment *z,* we can define the unit-level dose response function as

$$Y_i^{(z)}\ for\ z\,\mathrm{in}\,Z \tag{5.3}$$

And let the average dose-response function be

$$\mu(z) = E[Y_i^{(z)}] \tag{5.4}$$

Next define the multilevel GPS for continuous treatments as

$$E = e(Z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r}) \tag{5.5}$$

where

$$e(z, \vec{x}, \vec{w}, \vec{x}\vec{w}, \vec{x}\vec{r}) = f_{Z\mid\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(z \mid \vec{x}, \vec{w}, \vec{x}\vec{w}, \vec{x}\vec{r}) \tag{5.6}$$

and $e(z, \vec{x}, \vec{w}, \vec{x}\vec{w}, \vec{x}\vec{r})$ is the conditional density of the treatment given the covariates.

Theorem $$f_Z(z \mid e_i(z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r}), Y_i^{(z)}) = f_Z(z \mid e_i(z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r})) \tag{5.7}$$

For each *z* define a joint law for $(Y_i^{(z)}, \vec{Z}, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r}, e(z, \vec{X}, \vec{W}, \vec{X}\vec{W}, \vec{X}\vec{r})$. Let *F* denote a conditional probability distribution and *f* denote a conditional density. Then using iterated integrals as in Hirano & Imbens (2004)

$$
\begin{aligned}
f_Z(z \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})) &= \int f_Z(z \mid \vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r},e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}))dF_{\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(\vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r} \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})) \\
&= \int f_Z(z \mid \vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})dF_{\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(\vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r} \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})) \\
&= \int e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})dF_{\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(\vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r} \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})) \\
&= e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})
\end{aligned}
$$

Now using Hirano and Imbens (2004) the left hand side of the equation can be rewritten as

$$f_Z(z \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}),Y^{(z)}) = \int f_Z(z \mid \vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r},e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}))dF_{\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(\vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r} \mid Y^{(z)}e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}$$

and by the assumption of ignorability

$f_Z(z \mid \vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r},e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}),Y^{(z)}) = f_Z(z \mid (\vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r})$ then

$$
\begin{aligned}
f_Z(z \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}),Y^{(z)}) &= \int e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})dF_{\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}}(\vec{x},\vec{w},\vec{x}\vec{w},\vec{x}\vec{r} \mid Y^{(z)}e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})) \\
&= e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r})
\end{aligned}
$$

Therefore for each z, $f_Z(z \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}),Y^{(z)}) = f_Z(z \mid e_i(z,\vec{X},\vec{W},\vec{X}\vec{W},\vec{X}\vec{r}))$.

APPENDIX B

Sensitivity Analysis Results

Table(5.8): Sensitivity Analysis of Reading Comprehension Effect

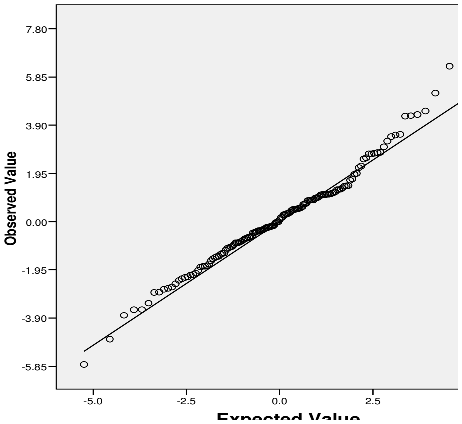
	Reading Comprehension			
Adjustment Variable	Adjusted Effect of TK ($\boldsymbol{\delta}^*$)	SE	T	P
Unadjusted (original effect of teacher Knowledge, $\boldsymbol{\delta}$)	0.096	0.036	2.689	0.008
Practice	0.106	0.036	2.958	0.003
Male	0.098	0.036	2.737	0.007
White Teacher	0.092	0.036	2.576	0.010
African-American Teacher	0.087	0.036	2.445	0.015
Hispanic Teacher	0.095	0.036	2.666	0.008
Asian Teacher	0.097	0.036	2.711	0.007
Bachelors in Elementary Education	0.094	0.036	2.639	0.009
Bachelors in Early Childhood Education	0.096	0.036	2.689	0.008
Bachelors in Literacy Education	0.095	0.036	2.670	0.008
Bachelors in Special Education	0.095	0.036	2.668	0.008
Masters Degree	0.096	0.036	2.690	0.008
Masters in Elementary Education	0.100	0.036	2.803	0.005
Masters in Early Childhood Education	0.096	0.036	2.682	0.008
Masters in Literacy Education	0.099	0.036	2.766	0.006
Masters in Special Education	0.096	0.036	2.688	0.008
Post Masters Degree	0.096	0.036	2.697	0.007
Standard Certification	0.091	0.036	2.555	0.011
Provisional Certification	0.096	0.036	2.697	0.007
Reading Certification	0.091	0.036	2.536	0.012
Special Education Certification	0.097	0.036	2.720	0.007
Number of Professional Trainings	0.081	0.036	2.262	0.024
High teaching experience (three+ years)	0.091	0.036	2.547	0.011
Teacher New to Reading First in 2006-07 School Year	0.090	0.036	2.529	0.012
Average age of students in class	0.095	0.036	2.659	0.008
Average non-sense word fluency score in class	0.044	0.036	1.234	0.218
Proportion male in class	0.088	0.036	2.469	0.014
Proportion in special education in class	0.096	0.036	2.685	0.008
Proportion eligible for free or reduced lunch in class	0.071	0.036	1.981	0.048
Proportion labeled as disabled	0.096	0.036	2.680	0.008
Proportion with limited English proficiency	0.096	0.036	2.673	0.008
Proportion Hispanic	0.098	0.036	2.729	0.007
Proportion Asian	0.095	0.036	2.661	0.008
Proportion White	0.074	0.036	2.065	0.040
Proportion African-American	0.069	0.036	1.940	0.053

Original standard error provides a reasonable estimate of the standard error of the adjusted estimate as by assumption U is unrelated to all other variables (Hong & Raudenbush, 2006)

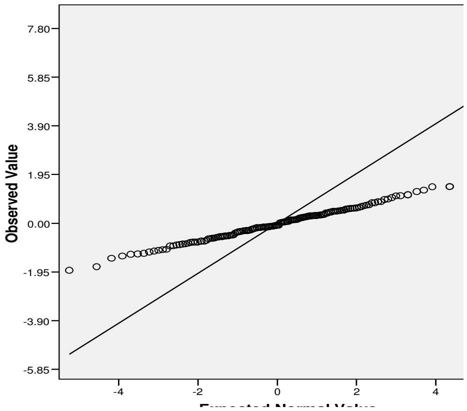
APPENDIX C

Additional Plots for Study 1

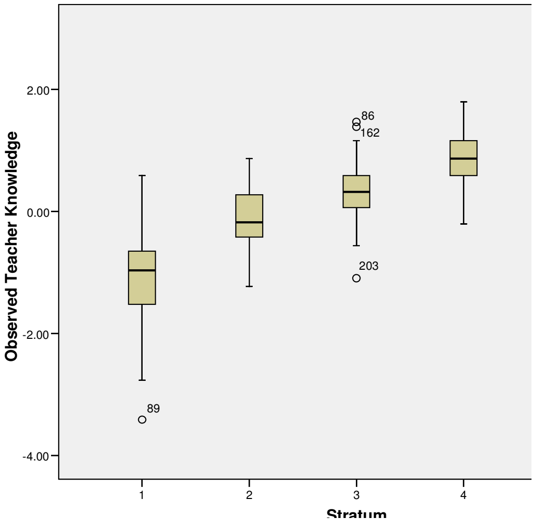
Figure(5.9): Quantile-Quantile Plot of Bi-variate t-values examining the association between teacher knowledge and covariates



Figure(5.10): Quantile-Quantile Plot of Tri-variate t-values examining the association between teacher knowledge and covariates controlling for the estimated PS



Figure(5.11): Variability in Observed Teacher Knowledge by Propensity Strata



APPENDIX D

List of Pretreatment Covariates

Taken from Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association*.

Student-level covariates

Demographic characteristics

1. Age at kindergarten entry (P1AGEENT)
2. Gender (FEMALE)
3. Hispanic (HISPANIC)

Assessment scores

4. C4 reading IRT scale score (C4RRSCAL)
5. C4 math IRT scale score (C4RMSCAL)
6. C1 reading IRT scale score (C1RSCALE)
7. C1 math IRT scale score (C1MSCALE)
8. C1 general knowledge IRT scale score (C1GSCALE)
9. C2 reading IRT scale score (C2RSCALE)
10. C2 math IRT scale score (C2MSCALE)
11. C2 general knowledge IRT scale score (C2GSCALE)

Home SES/Home language/Home size and structure

12. SES (WKSESL)
13. Poverty (WKPOVRTY)
14. Mother's education (MOMED)
15. Mother worked between child birth and kindergarten (MOMWK1)
16. English as home language (HOMELANG)
17. Number of siblings (P1NUMSIB)
18. Single parent family (SGPARENT)
19. Two-parent family (TWPARENT)

Home literacy environment and other activities

20. Amount of books at home (P1CHLBOO)
21. Fall, K parent report of child's frequency of reading books outside school (P1CHRBK1)
22. Spring, K parent report of child's frequency of reading books outside school (P2CHRBK1)

23. Home computer for child use (P2HOMECM)
24. Extracurricular lessons (P2EXTR)

Parent involvement in school and parenting

25. Current school chosen (P1SCHCHC)
26. Tuition paid for education (P2TUITIO)
27. Parent educational expectation (P1EXPCTN)
28. Parent attending school activities (P2NATTEN)
29. Parents' negative attitude toward parenting (NEGAPRT)

Day care/Preschool/Head Start learning experience/Special services for the child

30. Child ever in center-based care (P1CENTER)
31. Child with reduced/free lunch (P2FRELCH)
32. Child receiving special service/education (P2SPECND)

Child physical and mental health

33. Spring, K child fell behind due to health (T2FLBHND)
34. Child with disability (P1DISABL)

Teacher assessment of student status at the beginning and by the end of the kindergarten year

35. Fall, K child literacy ARS score (T1ARSLIT)
36. Fall, K child math ARS score (T1ARSMAT)
37. Fall, K child general knowledge ARS score (T1ARSGEN)
38. Fall, K teacher rating on child approaches to learning (T1LEARN)
39. Fall, K teacher rating on child self control (T1CONTRO)
40. Fall, K teacher rating on child interpersonal skills (T1INTERP)
41. Fall, K teacher rating on child externalizing problem behaviors (T1EXTERN)
42. Fall, K teacher rating on child internalizing problem behaviors (T1INTERN)
43. Spring, K child literacy ARS score (T2ARSLIT)
44. Spring, K child math ARS score (T2ARSMAT)
45. Spring, K child general knowledge ARS score (T2ARSGEN)
46. Spring, K teacher rating on child approaches to learning (T2LEARN)
47. Spring, K teacher rating on child self control (T2CONTRO)
48. Spring, K teacher rating on child interpersonal skills (T2INTERP)
49. Spring, K teacher rating on child externalizing problem behaviors (T2EXTERN)
50. Spring, K teacher rating on child internalizing problem behaviors (T2INTERN)
51. Spring, K teacher rating on child language skills (T2RTLANG)
52. Spring, K teacher rating on child science/social studies skills (T2RTSCI)
53. Spring, K teacher rating on child math skills (T2RTMTH)
54. Spring, K teacher report on child not working at best ability (T2ABIL)

Teacher report on instructional services for child

55. Spring, K teacher report on child receiving individual tutored reading (T2TTRRD)
56. Spring, K child in pull-out small group in reading (T2SGRDG)
57. Spring, K child receiving individual tutored mathematics (T2TTRMTH)
58. Spring, K child in pull-out small group in math (T2SGMTH)
59. Spring, K child in in-class ESL program (T2INCESL)
60. Spring, K child in gifted/talented program (T2GFTTAL)
61. Spring, K child in program for behavioral problems (T2BEHPRB)
62. Spring, K child in Title I reading (T2TT1RD)
63. Spring, K child being active in structured play (T2STRPLY)
64. Spring, K child being active in unstructured play (T2UNPLAY)
65. Spring, K number of achievement reading groups in class (T2NORDGP)
66. Spring, K child in the lowest reading group (T2RDGPLO)
67. Spring, K child moving to a higher reading group (T2GPMVHI)

Parent assessment of student status at the beginning and by the end of the kindergarten year

68. Fall, K parent rating on child approaches to learning (P1LEARN)
69. Fall, K parent rating on child self control (P1CONTRO)
70. Fall, K parent rating on child social interaction (P1SOCIAL)
71. Fall, K parent report on child being sad/lonely (P1SADLON)
72. Fall, K parent report on child being impulsive/overactive (P1IMPULS)
73. Spring, K parent rating on child approaches to learning (P2learn)
74. Spring, K parent rating on child self control (P2CONTRO)
75. Spring, K parent rating on child social interaction (P2SOCIAL)
76. Spring, K parent report of child being sad/lonely (P2SADLON)
77. Spring, K parent report of child being impulsive/overactive (P2IMPULS)

Teacher report on parent involvement in school

78. Spring, K teacher report on parent attending conferences (T2REGCON)
79. Spring, K teacher report on parent coming for informal meetings (T2INFMTS)
80. Spring, K teacher report on parent volunteering in school (T2VOLUNT)
81. Spring, K teacher report on having other communications with parent (T2TCHCNF)

Classroom-level covariates

Kindergarten learning experience

82. Full-day kindergarten in Fall, K (A1HLFDAY)
83. Full-day kindergarten in Spring, K (A2HLFDAY)
84. Number of class hours per day in Fall, K (A1HRSDA)
85. Days per week in Fall, K (A1DYSWK)
86. Number of class hours per week in Fall, K (A1TIMEWK)
87. Less than five days per week in Fall, K (A1SHRTWK)
88. Time on teacher-directed whole class activity in Fall, K (B1TDWCLS)
89. Time on reading and language arts instruction per day in Spring, K (A2RDLHR)
90. Content coverage in reading and language arts curriculum in Spring, K (RDCURR)
91. Encouraging invented spelling in Spring, K (A2INVSPE)
92. Instructional activities with math symbols in Spring, K (SYMBLMTH)
93. Content coverage in math curriculum in Spring, K (MTHCURR)
94. Teacher frequency of borrowing books from library in spring, K (A2BORROW)
95. Fall, K teacher report of unpaid preparation hours per week (B1NOPAYP)
96. Fall, K teacher having different standards based on talent

Kindergarten class composition/assignment

97. Percentage of Hispanics in Fall, K class (A1PHIS)
98. Proportion of 3 and 4 year olds in Fall, K class (A1PR34)
99. Proportion of boys in Fall, K class (A1PRBOYS)
100. Percentage of children in Fall, K class with preschool records (A1PCPRE)
101. Class enrollment in Fall, K (A1TOTAG)
102. Number of children in Fall, K class repeating kindergarten (A1REPK)
103. Proportion of children in Fall, K class repeating kindergarten (A1PRREPK)
104. Number of children in class recognizing letters at the start (A1LETTC)
105. Teacher rating of behavior from the teacher in fall, K (A1BEHVR)
106. Number of LEP students in Fall, K class (A1NMLEPC)
107. Proportion of LEP students in Fall, K class (A1PRLEP)
108. Number of students classified as gifted in Spring, K class (A2GIFT)
109. Number of students taking part in gifted/talented program in Spring, K class (A2PRTGF)
110. Number of students with disabilities in Spring, K class (A2DISAB)
111. Kindergarten assignment in Fall, K based on preschool experience (A1PRESC)

Kindergarten teacher characteristics

112. Fall, K teacher holding different standards based on students' capability (B1DIFSTD)
113. Fall, K teacher spending unpaid preparation hours per week (B1NOPAYP)
114. Fall, K teacher being Hispanic/Latino (B1HISP)
115. Spring, K teacher being Hispanic/Latino (B2HISP)

116. Fall, K teacher years of experience of teaching preschool (B1YRSPRE)
117. Spring, K teacher years of experience of teaching preschool (B2YRSPRE)
118. Fall, K teacher years of experience of teaching kindergarten (B1YRSKIN)
119. Spring, K teacher years of experience of teaching kindergarten (B2YRSKIN)
120. Fall, K teacher years of experience of teaching first grade (B1YRSFST)
121. Spring, K teacher years of experience of teaching first grade (B2YRSFST)
122. Fall, K teacher years of experience of teaching ESL (B1YRSESL)
123. Spring, K teacher years of experience of teaching ESL (B2YRSESL)
124. Spring, K teacher years of experience of teaching bilingual education (B2YRSBIL)
125. Fall, K teacher years of teaching experience at the school (B1YRSCH)
126. Spring, K teacher years of teaching experience at the school (B2YRSCH)
127. Fall, K teacher's educational degree beyond the Bachelor's (B1DEGREE)
128. Spring, K teacher's educational degree beyond the Bachelor's (B2DEGREE)
129. Fall, K teacher having taken ESL courses (B1ESL)
130. Spring, K teacher having taking ESL courses (B2ESL)
131. Fall, K teacher speaking other language (A1OTHLNG)

School-level covariates

School poverty level and student composition

132. Spring, K total school enrollment (S2KENRLS)
133. Spring, K percentage of minority students enrolled (S2KMINOR)
134. Spring, K school percentage of gifted/talented students (S2KGFTED)
135. Spring, K school percentage of Hispanic students (S2PCTHSP)

School type

136. School enrollment as of 10/1/1998 (S2ANUMCH)
137. Spring, K school instructional level (S2KSCLVL)
138. Spring, K school average daily attendance (S2ADA)
139. Spring, K school with ungraded classroom or transitional grade (S2UNGRAD, S2TRANS)
140. Spring, K school with Kindergarten (S2KINDER)
141. Spring, K school with grade levels between grade 1 – 5 (S2GRDLV1)
142. Spring, K school with grade levels beyond grade 5 (S2GRDLV6)
143. Spring, K school being public (S2PUBLIC)
144. Spring, K school enrollment requiring academic records (S2ACADRC)
145. Spring, K school Title 1 funding used for professional development (S2TT1PD)
146. Spring, K school Title 1 funding used for other Title 1 purposes (S2TT1OTH)
147. Spring, K school percentage of LEP children (S2LEPSCH)
148. Spring, K school percentage of LEP children in K, transitional K, and transitional first grade (S2LEPKND)
149. Spring, K special education students on IEP and 504 plan (S2IEP504)
150. Spring, K children with disability receiving special services (S2DSBNO)
151. Spring, K school with gifted/talented program in transitional K (S2GFTR)
152. Spring, K school with gifted/talented program in grade 4 and 5 (S2GFT4TH S2GFT5TH)
153. Spring, K school number of FTE classroom teachers (S2TCHFTE)
154. Spring, K school number of FTE bilingual-ESL teachers (S2ESLFTE)
155. Spring, K school lowest annual teacher salary (S2LOSLRY)
156. Spring, K school highest annual teacher salary (S2HISLRY)
157. Spring, K school percentage of Hispanic teachers (S2ETHNIC)
158. Spring, K school percentage of Asian teachers (S2Q62ASN)

School climate

159. Spring, K principal report of school emphasis on language and number skills goals (S2GOAL1)
160. Spring, K principal report of school emphasis on behavioral goals (BHVGOAL)
161. Spring, K principal report of school being successful in providing help to low achievers (S2SUCC7)
162. Spring, K principal report of school being successful in being open to new

ideas/methods (S2SUCC10)
163. Spring, K principal report of teacher union and administration working together (S2TOGTHR)
164. Spring, K principal report of standardized scores influencing evaluation of principal performance (S2SCORES)
165. Spring, K principal report of raising performance level of low-achieving students influencing evaluation of principal performance (S2PRFLVL)
166. Spring, K principal report of teacher and staff support influencing evaluation of principal performance (S2STFSPP)
167. Spring, K principal report of other factors influencing evaluation of principal performance (S2OTHINF)

Principal characteristics

168. Spring, K principal gender (S2GNDER)
169. Spring, K principal years of experience in teaching prek/Head start and K (S2YRPRKK)
170. Spring, K principal years of experience in teaching the 2nd grade or above (S2YR2ABV)
171. Spring, K number of courses principal having taken in early childhood education or child development(S2CRSECD)
172. Spring, K number of courses principal having taken having taken courses in ESL (S2CRSCDV)
173. Spring, K number of courses principal having taken having taken courses in administration (S2CRSADM)
174. Spring, K principal spending number of hours per week teaching (S2TEEECH)
175. Spring, K principal spending number of hours per week with required paperwork (S2PPRWRK)
176. Spring, K principal estimated percentage of children's names known (S2KNWNME)

School/teacher outreach to parent

177. Fall, K teacher report of emphasizing importance of home-assisted kindergarten learning (B1KLRN)
178. Fall, K parent report of school outreach to parents (P1OUTRCH)
179. Spring, K parent report of school outreach to parents (P2OUTRCH)
180. Spring, K teacher report of number of conferences with parents (A2NUMCNF)
181. Spring, K teacher report of parent often seeing child's work (A2SHARED)
182. Fall, K teacher report of giving parents orientation at the beginning of the year (B1PRNTOR)
183. Spring, K teacher report of giving parents orientation at the beginning of the year (B2PRNTOR)
184. Spring, K school report of offering orientation programs (S2ORIENT)
185. Spring, K teacher report of visiting students' homes before the beginning of the year (B2HMEVST).

186. Spring, K school report of frequency of PTA/PTO meetings (S2PTAMT)
187. Spring, K school report of frequency of sending report cards (S2RPRTCD)
188. Spring, K school report of frequency of parent-teacher conference (S2PTCONF)
189. Spring, K school report of frequency of performances for parents (S2INVITE)
190. Spring, K school report of frequency of classroom programs for parents (S2CLASPR)
191. Spring, K school report of frequency of fundraisers (S2FUNDRS)
192. Spring, K school report of assistance/outreach to LM-LEP families (S2LEPHLP)

Instructional resources

193. Spring, K teacher's use of instructional resources (A2RSRUSE)
194. Spring, K school accommodation size (S2CHLDNM)
195. Spring, K school number of instructional rooms (S2RMNUM)
196. Spring, K school number of instructional computers (S2INSTCM)
197. Spring, K teacher report of using computer equipment (A2COMP1)
198. Spring, K teacher report of using software (A2SOFTW1)
199. Spring, K teacher report of using heating/air conditioning (A2HTAC1)
200. Spring, K school adequacy of facility (S2FACLTY)

Neighborhood environment and school safety

201. Spring, K school having problem with gangs (S2GANGRC)
202. Spring, K children with weapons in school (S2WEAPON)
203. Spring, K visitors being required to sign in (S2SIGNIN)
204. Spring, K school taking more security measures (K2Q2SCRT)
205. Spring, K school safety rating (K2Q3)
206. Spring, K school with decorated hallways (K2Q6_A)

APPENDIX E

Covariates Used in Impact Based Propensity Scores

School type

School urbanicity
Percent female
SES
Mother's education
Length of school day
Percent with disability spring 1st year
Number classified as gifted or talented
Number of conferences with parent
Parent orientations
Average teacher experience in bilingual education
Math score fall 1st year
General score fall 1st year
Math score spring 1st year
Math score spring 1st year squared
Reading score spring of 1st year
Reading score spring of 1st year squared
Decorated hallways
Percent with disability fall 1st year
Parental approach to learning
Average number of siblings
Average self control of students
Average frequency parents laugh with kids
Average level of social interaction
Average daily attendance
Percent of Hispanic teachers
Frequency of fundraisers
Problems with gangs
Gifted/talented program in transitional K
Grade level Kindergarten
Percent of names known
School instructional level from SAQ
Percent with LEP students in K
Percent of LEP students in school
Percent of Hispanic students
Percent of Asian teachers
Frequency of report cards
Grade level transitional
Other title on purpose
Average teacher self control
Average teacher interpersonal skill
Average teacher approach to learning fall 1st year
Program for behavioral problems
Average teacher approach to learning spring 1st year
Average teacher rating of language skills
Average teacher rating of math skills
Average structured play comparison
Average tutored in math
Average tutored in reading
Average unstructured play comparison
White percent
Hispanic percent
Asian percent

Average number of teachers effect by health
Average problem internalizing problem behaviors

APPENDIX F

MSE and Threshold Derivations for Covariance Adjustment on the Propensity Score

$MSE_1 > MSE_2$

$bias_1^2 + Var_1 > bias_2^2 + Var_2$

$$(\frac{\rho_{yz}-\rho_{ye_1(x)}\rho_{ze_1(x)}}{1-\rho^2_{ze_1(x)}}-\theta)^2+(\frac{1-R_Y^2}{n-q-1})(\frac{1}{1-\rho^2_{ze_1(x)}}) > (\frac{\rho_{yz}-\rho_{ye_2(x)}\rho_{ze_2(x)}}{1-\rho^2_{ze_2(x)}}-\theta)^2+(\frac{1-R_Y^2}{n-q-1})(\frac{1}{1-\rho^2_{ze_2(x)}})$$

$Since\ R_Y^2 = 1-\frac{(\rho^2_{yz}+\rho^2_{ye_{.}(x)}-2\rho_{yz}\rho_{ye_{.}(x)}\rho_{ze_{.}(x)}}{1-\rho^2_{ze_{.}(x)}}$

$$(\frac{\rho_{yz}-\rho_{ye_1(x)}\rho_{ze_1(x)}}{1-\rho^2_{ze_1(x)}}-\theta)^2+(\frac{1-\rho^2_{ze_1(x)}-\rho^2_{yz}-\rho^2_{ye_1(x)}+2\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}}{(n-q-1)(1-\rho^2_{ze_1(x)})^2}) > (\frac{\rho_{yz}-\rho_{ye_2(x)}\rho_{ze_2(x)}}{1-\rho^2_{ze_2(x)}}-\theta)^2+(\frac{1-\rho^2_{ze_2(x)}-\rho^2_{yz}-\rho^2_{ye_2(x)}+2\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}}{(n-q-1)(1-\rho^2_{ze_2(x)})^2})$$

$$(\frac{\rho_{yz}-\rho_{ye_1(x)}\rho_{ze_1(x)}-\theta+\theta\rho^2_{ze_1(x)}}{1-\rho^2_{ze_1(x)}})^2+(\frac{1-\rho^2_{ze_1(x)}-\rho^2_{yz}-\rho^2_{ye_1(x)}+2\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}}{(n-q-1)(1-\rho^2_{ze_1(x)})^2}) > (\frac{\rho_{yz}-\rho_{ye_2(x)}\rho_{ze_2(x)}-\theta+\theta\rho^2_{ze_2(x)}}{1-\rho^2_{ze_2(x)}})^2+(\frac{1-\rho^2_{ze_2(x)}-\rho^2_{yz}-\rho^2_{ye_2(x)}+2\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}}{(n-q-1)(1-\rho^2_{ze_2(x)})^2})$$

$$(\frac{\rho^2_{yz}-2\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2\theta\rho_{yz}+2\theta\rho_{yz}\rho^2_{ze_1(x)}+\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+\theta^2-2\theta^2\rho^2_{ze_1(x)}+\theta^2\rho^4_{ze_1(x)}}{(1-\rho^2_{ze_1(x)})^2})+(\frac{1-\rho^2_{ze_1(x)}-\rho^2_{yz}-\rho^2_{ye_1(x)}+2\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}}{(n-q-1)(1-\rho^2_{ze_1(x)})^2}) >$$

$$(\frac{\rho^2_{yz}-2\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2\theta\rho_{yz}+2\theta\rho_{yz}\rho^2_{ze_2(x)}+\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+\theta^2-2\theta^2\rho^2_{ze_2(x)}+\theta^2\rho^4_{ze_2(x)}}{(1-\rho^2_{ze_2(x)})^2})+(\frac{1-\rho^2_{ze_2(x)}-\rho^2_{yz}-\rho^2_{ye_2(x)}+2\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}}{(n-q-1)(1-\rho^2_{ze_2(x)})^2})$$

$Let\ a = n-q-1 = DoF$

$$(\frac{a\rho^2_{yz}-2a\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2})+(\frac{1-\rho^2_{ze_1(x)}-\rho^2_{yz}-\rho^2_{ye_1(x)}+2\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) >$$

$$(\frac{a\rho^2_{yz}-2a\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2})+(\frac{1-\rho^2_{ze_2(x)}-\rho^2_{yz}-\rho^2_{ye_2(x)}+2\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2})$$

Combining bias and var

$$(\frac{(a-1)\rho^2_{yz}-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) >$$

$$(\frac{(a-1)\rho^2_{yz}-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2})$$

Setting equal to zero to find threshold

$$(\frac{(a-1)\rho^2_{yz}-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) \text{-}$$

$$(\frac{(a-1)\rho^2_{yz}-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) \mathbf{= 0}$$

Factoring out ρ^2_{yz}

$$(\frac{(a-1)\rho^2_{yz}}{a(1-\rho^2_{ze_1(x)})^2})+(\frac{-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2})$$

$$-(\frac{(a-1)\rho^2_{yz}}{a(1-\rho^2_{ze_2(x)})^2})-(\frac{-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) \mathbf{= 0}$$

$$(\frac{(a-1)\rho^2_{yz}}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)\rho^2_{yz}}{a(1-\rho^2_{ze_2(x)})^2})+(\frac{-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2})$$

$$-(\frac{-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) \mathbf{= 0}$$

$$\rho^2_{yz}\left[(\frac{(a-1)}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)}{a(1-\rho^2_{ze_2(x)})^2})\right]+(\frac{-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) \text{-}$$

$$-(\frac{-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) \mathbf{= 0}$$

Factoring out ρ_{yz}

$$\rho^2_{yz}\Big[(\frac{(a-1)}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)}{a(1-\rho^2_{ze_2(x)})^2})\Big]+(\frac{-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}+a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2})$$
$$-(\frac{-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}+a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) = 0$$

$$\rho^2_{yz}\Big[(\frac{(a-1)}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)}{a(1-\rho^2_{ze_2(x)})^2})\Big]+[\frac{-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}]+(\frac{a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) -$$
$$-[\frac{-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}]-(\frac{a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) = 0$$

$$\rho^2_{yz}\Big[(\frac{(a-1)}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)}{a(1-\rho^2_{ze_2(x)})^2})\Big]+[\frac{-2(a-1)\rho_{yz}\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}]-[\frac{-2(a-1)\rho_{yz}\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{yz}+2a\theta\rho_{yz}\rho^2_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}]$$
$$+(\frac{a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) -$$
$$(\frac{a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) = 0$$

$$\rho^2_{yz}\Big[(\frac{(a-1)}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)}{a(1-\rho^2_{ze_2(x)})^2})\Big]+\rho_{yz}\Big\{[\frac{-2(a-1)\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta+2a\theta\rho^2_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}]-[\frac{-2(a-1)\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta+2a\theta\rho^2_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}]\Big\}$$
$$+(\frac{a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) -$$
$$(\frac{a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) = 0$$

$$\rho^2_{yz}\Big[(\frac{(a-1)}{a(1-\rho^2_{ze_1(x)})^2})-(\frac{(a-1)}{a(1-\rho^2_{ze_2(x)})^2})\Big]+\rho_{yz}\Big\{[\frac{-2(a-1)\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta+2a\theta\rho^2_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}]-[\frac{-2(a-1)\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta+2a\theta\rho^2_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}]\Big\}$$
$$+(\frac{a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}) -$$
$$(\frac{a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}+a\theta^2-2a\theta^2\rho^2_{ze_2(x)}+a\theta^2\rho^4_{ze_2(x)}+1-\rho^2_{ze_2(x)}-\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}) = 0$$

To solve for threshold in terms of ρ_{yz} :

Let $\mathrm{a} = \frac{a-1}{a(1-\rho^2_{ze_1(x)})^2}+\frac{-a+1}{a(1-\rho^2_{ze_2(x)})^2}$

$$\mathrm{b} = \frac{-2a\rho_{ye_1(x)}\rho_{ze_1(x)}+2\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta+2a\theta\rho^2_{ze_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}+\frac{+2a\rho_{ye_1(x)}\rho_{ze_2(x)}-2\rho_{ye_1(x)}\rho_{ze_2(x)}+2a\theta-2a\theta\rho^2_{ze_2(x)}}{a(1-\rho^2_{ze_2(x)})^2}$$

$$\mathrm{c} = \frac{a\rho^2_{ye_1(x)}\rho^2_{ze_1(x)}+2a\theta\rho_{ye_1(x)}\rho_{ze_1(x)}-2a\theta\rho_{ye_1(x)}\rho^3_{ze_1(x)}+a\theta^2-2a\theta^2\rho^2_{ze_1(x)}+a\theta^2\rho^4_{ze_1(x)}+1-\rho^2_{ze_1(x)}-\rho^2_{ye_1(x)}}{a(1-\rho^2_{ze_1(x)})^2}+$$
$$\frac{-a\rho^2_{ye_2(x)}\rho^2_{ze_2(x)}-2a\theta\rho_{ye_2(x)}\rho_{ze_2(x)}+2a\theta\rho_{ye_2(x)}\rho^3_{ze_2(x)}-a\theta^2+2a\theta^2\rho^2_{ze_2(x)}-a\theta^2\rho^4_{ze_2(x)}-1+\rho^2_{ze_2(x)}+\rho^2_{ye_2(x)}}{a(1-\rho^2_{ze_2(x)})^2})$$

Using the quadratic formula

$$\rho_{yz} = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

APPENDIX G

Individual Paper Abstracts

The Roles of Multilevel Propensity Scores and Variable Selection in Multilevel Settings

Abstract

Despite increasing popularity of propensity score (PS) methods, literature is relatively scarce concerning the use and construction of multilevel PS's in estimating a treatment effect. In this study I examined several PS model types including a traditional logistic model and a multilevel logistic model in conjunction with a hierarchical linear outcome model in estimating the treatment effect in terms of bias and mean-squared error (MSE). I then examined variable selection for such models by illustrating how the choice of variables included in the PS model can affect bias, variance and MSE of an estimated treatment effect. The results suggest consistent but minimal gains from multilevel PS's and, further, that variable selection in the PS model can play a large role, both relative to model type and in an absolute sense. The method is applied to a study concerning the effect of teacher's literacy knowledge on first graders' reading achievement.

Variable Selection in Propensity Score Models for Multilevel Settings

Abstract

The variables included in the estimation of a propensity score (PS) have a strong influence on the properties of the corresponding treatment effect estimator. This study developed a method to construct PS's in a manner that attempts to minimize the mean-squared error (MSE) of the corresponding treatment effect estimator. In particular, using commonly available outcome proxies such as pretest measures, the method utilizes each covariate's relationship with the treatment and the outcome to construct a PS model that attempts to jointly minimize bias and variance. The study specifically focuses on multilevel observational studies in education that utilize the PS in conjunction with a multilevel outcome model to adjust for confounding variables. The method is applied to a study concerning the effect of school retention policies on the average math and reading achievement scores.

Robustness of Causal Inferences in Binomial Regression Models

Abstract

Binomial regression models (BRMs) used to test a hypothesis concerning a treatment assume the analyst has included all confounding variables. However, in observational studies it is frequently difficult to identify and measure exhaustively all confounding variables potentially leading to false inferences. In this study, I developed an index to assess the sensitivity of inferences in observational studies by extending Frank's (2000) impact threshold of a confounding variable index from the linear model to BRMs. The extension is developed for both the simple BRM as well as the multiple BRM and is applied to a study concerning reading achievement in Limpopo, South Africa.

BIBLIOGRAPHY

Agresti, A., (1996). Categorical Data Analysis. New York, Wiley.

Anderson, S., Auquier, A., Hauck, W., Cakes, D., Vandaele, W., Weisberg, H., Bryk, A., & Kleinman, J., (1980). Statistical Methods for Comparative Studies. New York, Wiley.

Assessment Committee. (2002). Analysis of Reading Assessment Measures, Coding Form for Dynamic Indicators of Basic Early Literacy Skills. Downloaded from http:DIBELS.uoregon.edu in February 2002.

Ballou, D., (1996). Do public schools hire the best applicants? The Quarterly Journal of Economics 111, 1, pp.97–133. Becker, H. J. 1999

Baker, D., Goesling, B., & Letendre, G., (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley" effect on mathematics and science achievement. Comparative Education Review, vol. 46, No. 3, pp. 291-312.

Berkson, J., (1957). Tables for the Maximum Likelihood Estimate of the Logistic Function. Biometrics, vol. 13, 1, pp. 28-34.

Bloom, H. S. (2005). Randomized groups to evaluate placed-based programs, in learning more from social experiments: Evolving analytic approaches. New York: Russell Sage Foundation.

Bos, C. S., Mather, N., Narr, R. F., & Babur, N. (1999). Interactive, collaborative professional development in early literacy instruction: Supporting the balancing act. Learning Disabilities Research and Practice, 14, 4, pp. 227-238.

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., & Sturmer, T., (2006). Variable selection for PS models. Practice of Epidemiology, vol. 163, 12, pp. 1149-1156.

Buchmann, C., (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In A.C. Porter & A. Gamoran (Eds.), Methodological advances in cross-national surveys of educational achievement (pp. 150-197. Washington, DC: National Research Council, National Academy Press.

California Legislative Analyst's Office. (1997). Class size reduction Policy Brief. Sacramento, CA: California Legislative Analyst's Office, February 12, 1997. (http://www.lao.ca.gov/class_size_297.html)

Carlisle, J., Correnti, R., Phelps, G., Zeng, J. (In press). Exploration of the contribution of teachers' knowledge about reading to their students' improvement in reading. Reading and Writing: An interdisciplinary Journal.

Casella, G. & Berger, R., (2002). Statistical Inference. The Wadsworth Group, Pacific Grove, CA.

Cirino, P., Pollard-Durodola, S., Foorman, B., Carlson, C. & Francis, D., (2007). Teacher Characteristics, Classroom Instruction, and Student Literacy and Language Outcomes in Bilingual Kindergarteners. The Elementary School Journal, vol. 107, 4, pp. 341-364.

Cochran, W., (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, vol. 24, pp. 295-313.

Coe, M. & Hanita, M., (2003). Intraclass Correlation Values for Student Achievement Tests in Oregon. Paper presented at the 2009 Society for Research on Educational Effectiveness annual conference in Washington D.C.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. EQUALITY OF EDUCATIONAL OPPORTUNITY. Washington, D.C.: U.S. Government Printing Office.

Cook, T. D. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them, Educational Evaluation and Policy Analysis, 24, pp. 175-199

Correnti, R., & Rowan, B., (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. American Educational Research Journal, Vol. 44, No. 2, pp.298-338.

Cox, D., (1958). The Planning of Experiments. Wiley, New York.

Croninger, R.G., Rice, J.K., Rathbun, A., & Nishio, M. (2003). Teacher qualifications and first grade achievement. College Park, MD: Center for Education Policy and Leadership.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. Education Policy Analysis Archives, 8(1), 50.
Ehrenberg, R. G., & Brewer, D. J. (1995). Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. Economics of Education Review, 14, 1–21.

Featherman, D., & Hauser, R., (1976). Sexual Inequalities and Socioeconomic Achievement in the U.S. 1962-1973. American Sociological Review, 41, pp 462-483.

Foorman, B. R., & Moats, L. C. (2004). Conditions for sustaining research-based practices in early reading instruction. Remedial and Special Education, 25, 51–60.

Frank, K., (2000). Impact of a confounding variable on a regression coefficient. Sociological Methods and Research, vol. 29, 2, pp. 147-194.

Frank, K., Maurolis, S., Duong & Kelcey, B., (2009). Inferences from Randomized and Quasi- Experiments. Paper presented at the Society for Research on Educational Effectiveness 2009 annual conference in Washington D.C.

Garthwaite, P., Jolliffe, I. & Jones, B., (2002). Statistical Inference. Oxford Science Publications, New York.

Gastwirth, J., Krieger, A., & Rosenbaum, P., (1998). Dual and simultaneous sensitivity analysis for matched pairs. Biometrika, vol. 85, 4, pp. 907-920.

Gelman, A., & Meng, X., (2005). Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley, New York.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. Educational Evaluation and Policy Analysis, 22, 129–145.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. London: Sage Publications.

Hansen, B. B. (2004), Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association, 99, 609–618.

Hansen, B.B., & Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. Journal of Computational and Graphical Statistics,15, 609-627.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The market for teacher quality (Working Paper No. 11154). Cambridge, MA: National Bureau of Economic Research.

Hastie, T. J. and Pregibon, D. (1992) Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Haveman, R., & Wolfe, B. (1994). Succeeding generations: On the effects of investments in children. New York: Russell Sage Foundation.

Heckman, J. (2005). The Scientific Model of Causality. Sociological Methodology, 35, pp. 1-99.

Hirano, K., & Imbens, G., (2002). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services & Outcomes Research Methodology, 2, pp. 259–278.

Holland, P., (1986). Statistics and Causal Inference. Journal of the American Statistical Associations, Vol. 81, No.396, p. 947.

Holmes, C., (1989). Grade level retention effects: A meta analysis of research studies. In L.A. Shepard & M. L. Smith (Eds.), Flunking Grades: Research and policies on retention, pp. 64-78. London: The Falmer Press.

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

Hoover, H.D., Dunbar, S.B., & Frisbe, D.A. (2003). The Iowa tests: Norms and score conversions. Ithasca, IL: Riverside Publishing Co.

Imai, K., & Van Dyk, D., 2004. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association; Sep 2004; 99, 467; ABI/INFORM Global pg. 854

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

Kelcey, B., Rowan, B., Carlisle, J., & Phelps, G., (2008). The Effect of Teacher Knowledge on Student Achievement. University of Michigan Technical Report for Assessment of Reading Knowledge.

Kainz, K. & Vernon-Feagans, L. (2007). The ecology of early reading development for children in poverty. The Elementary School Journal, 107, 407-427.

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

Kleyman, Y., & Hansen, B., (2008). Test for Covariate Balance in Comparative Studies. Paper presented at the Joint Statistical Meetings 2008 annual conference.

Krueger, A., (1999). Experimental estimates of education production functions. Quarterly Journal of Economics, vol. 114, No. 2, pp. 497-532.

Lee, V., Zuze, T., & Ross, K., (2005). School Effectiveness in 14 Sub-Saharan African Countries: Links with 6th Graders' Reading Achievement. Studies in Educational Evaluation, 31, pp. 207-246.

Maas, C., & Hox, J., (2005). Robustness of multilevel parameter estimates against small sample sizes. Technical report Utrecht University.

McCullagh, P. & Nelder, J., (1983). Generalized Linear Models. Chapman and Hall: London.

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., Quiroga, T., & Gray, A. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. Journal of Learning Disabilities, 35, 69-86.

McCutchen, D., Harry, D.R., Cunningham, A.E., Cox. S., Sidman, S., & Covill, A.E. (2002). Reading teachers' knowledge of children's literature and English phonology. Annals of Dyslexia, 52, 207-228.

Mislevy, R. & Bock, R. (1997). Bilog: Item analysis and test scoring with binary models- Lincolnwood, IL: Scientific Software International, 1997

Moats, L.C. (2003). Language Essentials for Teachers of Reading and Spelling. Longmont, CO: Sopris West.

Moinedden, R., Matheson, F., & Glazier, R., (2007). A simulation study of sample size for multilevel logistic regression models. BMC Medical Research Methodology, pp. 7-34.

Morrison, F., Griffith, E. & Alberts, D., (1997). Nature-nuture in the classroom: Entrance age, school readiness and learning in children. Developmental Psychology, 33, 2, pp. 254-262.

NCLB, (2001). The No Child Left Behind Act of 2001. Public Law PL 107-110.

Nguyen, K., Wu, M. & Gillis, S. (2005). Factors Influencing Achievement Levels in SACMEQ II: Botswana: An Application of Structural Equation Modeling. Paper presented to the International Invitational Educational Policy Research Conference in Paris, France, September 28-October 2, 200

Pan, W., & Frank, K., (2004). An approximation to the distribution of the product of two dependent correlation coefficients. Journal of Statistical Computation and Simulation, vol. 74, No. 6, pp. 419-443.

Peugh, J.L. & Enders, C.K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. Review of Educational Research; Winter 2004; 74, 4; Research Library pg. 525.

Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teachers of reading. Elementary School Journal, 105, 31-48.

Pinheiro, J., & Bates, D., (2000). Mixed Effects Models in S and S-Plus. Springer, New York.

Raghunathan, T. E. Lepkowski, J. M. Van Hoewyk, J. Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey methodology 2001, VOL 27; PART 1, pages 85-96.

Rao, C., (1952). Advanced Statistical Methods in Biometric Research. Wiley and Sons, Inc, New York.

Raudenbush, S., & Bryk, A., (2002). Hierarchical Linear Models. Sage publishing, Thousand Oaks, CA.

Raudenbush, Stephen. (2005). "Learning from attempts to improve schooling: The contribution of methodological diversity." Educational Researcher, Vol. 34(5), 25-31, 2005.

Robins, J., Hernán, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11(5), 550–560.

Robins, J., Mark, S., & Newey, W., (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. Biometrics, vol. 48, pp. 479-495.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. JASA, 90, 122–129.

Robins, J.M., A. Rotnizky and L.P. Zhao (1995). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89, 846-866.

Roderick, M., Bryk, A., Jacobs, B., Easton, J., & Allensworth, E., (1999).Ending social promotion: Results from the first two years. Chicago: Chicago Consortium on School Research.

Rosenbaum, P. & Rubin, D., (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, vol. 79, 387, pp.516-524.

Rosenbaum, P., (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. Journal of the American Statistical Association, vol. 79, pp.41-48.

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

Rosenbaum, P., (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika, vol 74,1, pp.13-26.

Rosenbaum, P., (1988). Sensitivity analysis for matching with multiple controls. Biometrika, vol. 75, pp. 577-581.

Rosenbaum, P., (1989). Sensitivity analysis for matched observational studies with many ordered treatments. Scandinavian Journal of Statistics, vol. 16, pp. 227-236.

Rosenbaum, P. (1991). A Characterization of optimal designs for observational studies. Journal of the Royal Statistical Society, Series B, 53, 597–610.

Rosenbaum, P.R. (1993). Confident search. Journal of Computational and Graphical Statistics, 2, 381-403.

Rosenbaum, P., (1995). Observational Studies. Springer-Verlag, New York.

Rosenbaum. P. R., and Rubin, D. B. (1983a). The Central Role of the Propensity Score in Observational Studies for Casual Effects, Biometrika, 70, 41-55.

Ross, K., Saito, M., Dolata, S., Ikeda, M., and Zuze, L. (2004). Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) Data Archive for the SACMEQ1 and SACMEQ II Projects Version 4.0 (October, 2004) User s Guide. Paris: UNESCO.

Rubin, D., (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology, Vol. 66, No.5, p. 688-701.
Rubin, D., (1976). Inference and missing data. Biometrika, 63, 3, pp. 581-592.

Rubin, D. B. (1986), Which ifs have causal answers? Discussion of Holland's "Statistics and causal inference.". Journal of American Statistical Association, 83, 396.

Rubin, D. B. (1990), Formal Modes of Statistical Inference for Causal Effects. Journal of Statistical Planning and Inference, 25, 279_292.

Rubin, D., (1997). Estimating causal effects from large data sets using the propensity score. Annals of Internal Medicine, vol. 127, pp. 757-763.

Rubin, D., (2007). The design versus the analysis of observational studies for causal effects: Parallels
with the design of randomized trials. Stat. Med. 26 20–30.

Rubin, D. & Thomas, N., (1996). Matching using estimated propensity score: relating theory to practice. Biometrics, vol. 52, pp 249-64.

Scott, D. W. (1979) On optimal and data-based histograms. Biometrika 66, 605–610.

Seltzer, M., Kim, J., & Frank, K., (2006). Studying the sensitivity of inferences to possible unmeasured confounding variables in multisite evaluations. CSE technical report 701, CRESST, University of California, Los Angles.

Shepard, L., (1989). A review of research on kindergarten retention. In L.A. Shepard & M. L. Smith (Eds.), Flunking Grades: Research and policies on retention, pp. 64-78. London: The Falmer Press.

Shepard, L., & Smith, M., (1988). Academic and emotional effects of kindergarten retention on one school district. In L, A, Shepard & M.L. Smith (Eds.), Flunking Grades: Research and policies on retention, pp. 64-78. London: The Falmer Press.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4–14.

Simpson, G. A., & Fowler, M. G. (1994). Geographic mobility and children's emotional/behavioral adjustment and school functioning. Pediatrics, 93(2), 303-309.

Sobel, M., (1998). Causal inference in statistical models of process of socioeconomic achievement: A case study. Sociological Methods and Research, 27, pp318-348.

Spybrook, J., Raudenbush, S., Liu, X., & Congdon, R., (2006). Optimal Design for Longitudinal and Multilevel Research: Documentation for the “Optimal Design” Software.

The University of Iowa, (2003). Hoover, H.D., Dunbar, S.B., Frisbee, D.A., et al. (2003). Iowa Test of Basic Skills: Guide to research and development. Ithaca, ILL: Riverside Publishing.

Tucker, C. J., Marx, J., & Long, L. (1998). "Moving on": Residential mobility and children's school lives. Sociology of Education, 71(2), 111-129.

U.S. General Accounting Office. (1994). Elementary school children: Many change schools frequently, harming their education. Washington, DC: Author. ED 369 526.
Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* New York: Springer (4th ed).

Verbitsky, N., & Raudenbush, S., (2004). Causal inference in spatial settings. Proceedings of the American Statistical Association, Social Statistics Section [CD-ROM], American Statistical Association, 2369-2374, Alexandria, VA.

Wood, D., Halfon, N., Scarlata, D., Newacheck, P., & Nessim, S. (1993). Impact of family relocation on children's growth, development, school function, and behavior. Journal of the American Medical Association, 270(11), 1334-1338.

Weiler, K, & Mitchell, C (1992), What schools can do: Critical pedagogy and practice (pp.177-202). Buffalo: State University of New York Press

Winship, C., & Morgan, S. (1999). The Estimation of Causal Effects from Observational Data. Annual Review of Sociology, 25, 659_707.