A BAYESIAN PROPENSITY SCORE APPROACH FOR MULTILEVEL OBSERVATIONAL STUDIES

by

Jianshen Chen

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Educational Psychology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2014

Date of final oral examination: 08/12/2014

The dissertation is approved by the following members of the Final Oral Committee:
David Kaplan, Professor, Committee Chair, Department of Educational Psychology
Jee-Seon Kim, Professor, Department of Educational Psychology
James Wollack, Associate Professor, Department of Educational Psychology
Peter Steiner, Assistant Professor, Department of Educational Psychology
Geoffrey Borman, Professor, Department of Educational Leadership & Policy Analysis

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Professor David Kaplan with all my heart for his invaluable support, guidance and love throughout my Ph.D. study. Without him, I would not be the person I am today. He led me into the exciting world of Bayesian inference and propensity score methods. I conducted a series of research on the topic of Bayesian propensity score approaches with him, which inspired me to have the idea of this dissertation to extend the two-step Bayesian propensity score approach we developed into the multilevel settings.

I owe my gratitude to Professor Jim Wollack, who was my supervisor at the Testing & Evaluation (T&E) Services at the University of Wisconsin-Madison (UW-Madison). Jim greatly broadened my knowledge and experience in psychometrics through his supervision of my work at T&E and an advanced graduate seminar he taught. He is always happy to share with me his abundant experience in research and life from a friend's perspective. Talking with him is a great joy.

I am grateful to Professor Peter Steiner, who taught me the state-of-art methods in quasi-experiments and often generously shared his research ideas and insightful advice of my work with me. Particularly, his constructive comments helped me greatly improve this dissertation.

I am indebted to Professor Jee-Seon Kim and Daniel Bolt, who have not only taught me professional knowledge in quantitative methods such as experimental design, multilevel modeling and classical and modern test theory, but also provided me precious support and warmth like family members through my graduate life.

I am thankful to Professor Geoffrey Borman for his service in my committee and helpful advice on my dissertation.

I want to thank Professor Guanglei Hong at the University of Chicago for sharing the data set used in one of her papers with me, which I adopted as the real data for the case study in this dissertation.

I would like to thank my colleague Courtney Hall and UW-Madison research computing facilitator Lauren Michael for their kind help with regard to the cloud computing tool HTCondor, which greatly reduced the computing time of the

simulation studies in this dissertation.

I also wish to acknowledge all my friends in Madison, the professors, staff members and fellow graduate students in the Department of Educational Psychology at the UW-Madison, who have helped me a lot in various ways and made my study and life colorful and enjoyable. In particular, I would like to thank my friends Yi Lu, Sien Deng, Seo Young Lee and Soojin Park for the consistent help and love they gave me.

In addition, I acknowledge the funding support from the Grant R305D110001 of the Institute of Education Sciences, U.S. Department of Education, awarded to the University of Wisconsin-Madison (David Kaplan, PI). The opinions expressed here, though, are those of the author and do not necessarily represent views of the Institute or the U.S. Department of Education.

Last but not the least, I am most indebted to my husband Qi Tang and my parents for their unconditional love and support as I pursued this degree. I would also love to thank my ten-month old son Nathan C. Tang, who amazes me every day and brings me enormous joy. I dedicate this dissertation to them.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Propensity score technique is an effective tool to reduce selection bias in observational studies where the treatment selection process is not random. Multilevel data structure is commonly seen in educational and social science research, such as students nested in schools and individuals nested in social groups. Ignoring the multilevel structure leads to underestimated variance and inflated type I error rate. The extension of propensity score analysis into the multilevel settings can account for both selection bias and dependencies within clusters for nonrandomized nested trials. However, due to the two-stage nature of the propensity score procedure, the uncertainty of the propensity score is hard to be incorporated within the frequentist framework and is even more complicated in the multilevel context. A Bayesian perspective to propensity score analysis is promising in terms of uncertainty estimation by virtue of the ability to naturally incorporate parameter uncertainty. Also, a Bayesian propensity score approach allows the elicitation of prior information to enhance causal inference. Therefore, this dissertation develops a two-step Bayesian propensity score approach for multilevel observational studies and examines its properties through two comprehensive simulation studies and a real-data case study. Specifically, the effects of different sample sizes, priors, intra-class correlations, model specifications, matching strategies and propensity score methods on the treatment effect and variance estimation are evaluated. Results of the simulation studies show that the proposed approach offers less biased treatment effect estimate and more accurate uncertainty estimate compared to models that ignore the multilevel structure. Results of the simulation studies also indicate that Bayesian random intercept and slope propensity score model with optimal full matching via within-cluster matching is recommended when the within-cluster sample size is sufficient to facilitate close matches. Results of the case study reveal a practical limitation of the within-cluster matching strategy when the within-cluster sample size is small. Across-cluster matching might be used as an alternative when the cluster-level covariates are well controlled.

1 INTRODUCTION

1.1 Research background

It is well established that in randomized experiments, individuals are assigned to treatment conditions on the basis of a known probabilistic mechanism. By contrast, in nonrandomized studies, individuals are observed to participate in one or another treatment conditions, where the treatment selection mechanism is unknown. The nonrandom process of treatment selection often introduces selection bias that may result in highly unbalanced covariates in different treatment conditions and thus weakens causal inference. For example, if the causal effect of interest is the grade retention effect in the first grade on students' later academic performance for assisting with setting a retention policy, it is likely that retained students on average have lower family social economic status (SES) compared to promoted students so that the effects of SES on children's academic achievement could confound with the casual effect of interest. That is, students' social economic status can distribute distinctly for retained children compared to promoted children such that the direct comparison of their achievement scores is unfair if not based on similar family SES. Thus, the resulting estimated causal effect is biased.

Rosenbaum and Rubin (1983) proposed propensity score analysis (PSA) as an effective tool for reducing selection bias in observational studies through balancing on measured covariates. Since then, a large amount of literature has emerged with regard to both the estimation and the application of the propensity score. Models for estimating the propensity score have included, for example, parametric logit regression with chosen interaction and polynomial terms (see Dehejia and Wahba, 1999; Hirano and Imbens, 2001), and data mining techniques such as generalized boosted model (McCaffrey et al., 2004) and classification and regression trees (Lee et al., 2009). The most common methods for implementing the propensity score to estimate the treatment effect include stratification (Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004), weighting (Hirano and Imbens, 2001; Lee et al., 2010), matching (Hansen, 2004; Rosenbaum, 1989) and regression adjustment (Ru-

bin, 1979). Details about these approaches are discussed later in section 1.3. All these propensity score techniques have a common goal, which is to achieve balanced covariate distributions across the different treatment conditions. Covariate balance after propensity score adjustment, therefore, is often evaluated to assess the effectiveness of propensity score techniques (see e.g., Harder et al., 2010).

Propensity score analysis and methods have been mainly developed within the frequentist framework of statistics. As another school of statistics that is distinct from frequentist, Bayesian inference views parameters as random and estimates them through a pre-specified probability distribution together with observed data. It has received an increasing popularity in many areas such as education, psychology, social science, epidemiology, business etc., as a result of its ability of incorporating prior information, naturally modeling uncertainty and intuitive probabilistic interpretation. A Bayesian perspective for propensity score analysis was originally advocated by Rubin (1985). However, the implementation of Bayesian approach to propensity score analysis was not illustrated. As one of the very few papers that studied Bayesian propensity score analysis, McCandless et al. (2009) provided the first practical Bayesian propensity score approach that stratifies individuals on the quintiles of the estimated propensity score, treating the propensity score as a latent variable and jointly estimating the propensity score and outcome models. Following McCandless et al. (2009)'s study, An (2010) presented a Bayesian approach for propensity score regression adjustment and matching methods that also jointly models the propensity score equation and outcome equation in one step.

A concern surrounding the work of McCandless et al. (2009) and An (2010) centers on the joint modeling of the propensity score equation and outcome equation. Specifically, the propensity score should only be determined by covariates measured prior to treatment implementation and not influenced by the outcome measured after treatment selection (Steiner and Cook, 2013). To address this problem, McCandless et al. (2010) utilized an approximate Bayesian technique for cutting undesirable feedback between propensity score model and outcome model components, but it often failed to converge. Chib and Greenberg (2010) described the

idea of modeling propensity score model and outcome model within the Bayesian framework at two steps parametrically and semi-parametrically but few details in regard to analytic procedures or estimators were provided to guide the statistical inference. Most recently, Kaplan and Chen (2012) outlined a complete framework of a two-step Bayesian propensity score approach, elaborated the treatment effect and variance estimators and examined its performance via propensity score stratification, weighting, and optimal full matching methods.

An extension of propensity score analysis is to apply it into the multilevel settings. Multilevel/hierarchical data structure is very common in a variety of areas (e.g., education, psychology, social science and public health), where individuals are typically nested within different sites or organization units, such as students nested in schools, patients nested in hospitals or clients nested in clinics. A special type of hierarchical structure can be seen in longitudinal studies, in which repeated measures across time are nested in individuals. The observations within each cluster share the same cluster characteristics and are often dependent of each other, such that ignoring the hierarchical structure or cluster effects can lead to seriously underestimated variation and thus distorted inference.

Multilevel modeling has been developed to account for cluster effects when data has hierarchical structure and when cluster-level units can be viewed as a random sample of some population. The multilevel technique was built upon a series of statistical development since 1960s (e.g., Elston and Grizzle, 1962; Rao, 1972; Lindley and Smith, 1972) and has received increasing attentions and applications in a broad range of areas. Until now, it has been widely used to study associative relationships for studies with nested structure and to investigate causal relationships in hierarchical randomized trials. In a well-implemented cluster randomized experiment, the treatment assignment is random and not affected by cluster characteristics. Therefore, only the hierarchical structure of the potential outcomes needs to be taken into account for model specification. However, in multilevel observational studies, in addition to possible dependencies among potential outcomes within each cluster, the treatment selections are often affected by both individual-level and cluster-level characteristics and vary across clusters. For instance, a child's

choice of going to full-day versus half-day kindergarten may depend on the district he or she lives in and the selections of other children who live nearby. Or take an example of students who are nested in schools, a student's selection of attending an education program may affect the choices of other students in the same school such that students' choices are more similar within each school compared to the selection of students from other schools.

Despite that observational studies with nested data structure are commonly seen, only a very limited amount of studies explored the use of multilevel modeling in nonrandomized designs for causal inference (see e.g., Hong and Raudenbush, 2006; Arpino and Mealli, 2011; Thoemmes and West, 2011). In addition, due to the two-stage nature of the propensity score procedure, the uncertainty of the propensity score is hard to be incorporated within the frequentist framework and is even more complicated in the multilevel context. A Bayesian perspective to propensity score analysis is promising to solve this problem by virtue of the ability to naturally incorporate parameter uncertainty. Also, Bayesian propensity score approach allows the elicitation of prior information to enhance the causal inference.

A review of literature reveals that a Bayesian propensity score approach for multilevel observational studies has not been studied yet. Therefore, this dissertation develops a two-step Bayesian propensity score approach for multilevel observational studies and examines its properties through two comprehensive simulation studies and a real-data case study. Specifically, in the simulation studies, the effects of different level-one and level-two sample sizes, priors, matching strategies, model specifications and intra-class correlations on the treatment effect and uncertainty estimation are evaluated via optimal full matching and regression adjustment methods.

The dissertation is organized as follows. In the rest parts of Chapter 1, the framework of the Neyman-Rubin causal model, i.e., potential outcome framework, is first introduced, based on which propensity score methods are defined. Then propensity score methods as well as their Bayesian counterparts are reviewed. Particularly, the two-step Bayesian propensity score approach developed by Kaplan and Chen (2012) is elaborated. In Chapter 2, the propensity score methods in

the multilevel settings are discussed. The proposed two-step Bayesian multilevel propensity score approach is then outlined in Chapter 3. This is then followed by the design and results of two comprehensive simulation studies and one real-data case study in Chapter 4 and 5, respectively. The dissertation closes with discussion and the implementation of the findings in Chapter 6.

1.2 The Neyman-Rubin Causal Model

For this dissertation, I follow the general notation of the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974). To begin, let T be a treatment indicator, representing the treatment selection. In the case of a binary treatment $T = \{0, 1\}$, for individual i, $T_i = 1$ if the individual received the treatment, and $T_i = 0$ if the individual did not receive the treatment. For individual i, the goal, ideally, would be to observe the individual under receipt of the treatment and under non-receipt of the treatment so that the causal effect at the individual level can be obtained. More formally, the *potential outcomes* framework for causal inference can be expressed as

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i} \tag{1.1}$$

where Y_i is the potential outcome of interest for person i, Y_{1i} is the potential outcome for individual i when exposed to the treatment, and Y_{0i} is the potential outcome for individual i when not exposed to the treatment. However, as Holland (1986) points out, the potential outcomes framework encounters a difficulty, which is known as *fundamental problem of causal inference*. That is, it is impossible to observe the value of Y_0 and Y_1 on the same individual i, and therefore impossible to observe the effect of Y_{1i} and Y_{0i} simultaneously.

The statistical solution to the *fundamental problem* is to utilize the information from the population of the individuals (e.g., Rubin, 1974; Rosenbaum and Rubin, 1983). In this case, two causal estimands may be of interest. The first is the *average*

treatment effect, ATE, defined as

$$\gamma_{ATE} = E(Y_{1i}) - E(Y_{0i}). \tag{1.2}$$

When the potential outcomes can be assumed independent from the treatment selection, the average treatment effect is equivalent to the *prima facie* effect, which is $E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 0)$. Returning to the grade-retention example, the *prima facie* effect compares the effect of, say, being retained in the first grade for those retained children, i.e. $E(Y_{1i}|T_i = 1)$, to the effect of not being retained in the first grade among those who, in fact, were not retained and promoted to the next grade instead, i.e., $E(Y_{0i}|T_i = 0)$. However, as pointed out by Heckman (2005), the policy or clinical question often lies in comparing $E(Y_{1i}|T_i = 1)$ to the counterfactual group $E(Y_{0i}|T_i = 1)$. In this case, the causal estimand of interest is the *average treatment effect on the treated, ATT*, defined as

$$\gamma_{ATT} = E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 1). \tag{1.3}$$

The essential difference between the ATE and ATT estimands is that for the ATE estimand it is assumed that a unit is drawn from a population and assigned either to the treatment group or the control group. By contrast, the ATT estimand assumes that an individual drawn from a population is assigned to the treatment group and the question concerns the outcome of that individual had he/she been assigned to the control group.

1.3 Propensity Score Analysis and Its implementation

Propensity score methods have been widely used to reduce selection bias while estimating causal effects in observational studies or randomized studies that suffer from attrition and/or treatment noncompliance (Barnard et al., 2003). Recently, propensity score matching has also been applied to remove remaining selection

bias in the cluster randomized trials when there are many confounding variables but sample size is small such that a small amount of stratification can not balance all the important covariates simultaneously (Xu and Kalbfleisch, 2010).

Throughout this dissertation, I will consider a simple case of a two-group problem, where individuals self-select into a "treatment group" or a "control group". The propensity score is then defined to be the conditional probability that a participant selected into the treatment group given a vector of covariates measured prior to the treatment selection. Following the notation of the Neyman-Rubin causal model, the single-level propensity score of individual i given a vector of observed covariates X_i can be expressed as $e(X_i) = P(T_i = 1|X_i)$.

All propensity score techniques rely on the assumption of *strong ignorability* (Rosenbaum and Rubin, 1983), which states that the potential outcomes are assumed to be independent of treatment assignment given observed covariates. If strong ignorability does not hold, then there may be hidden biases due to unobserved covariates. Under strong ignorability, participants with the same propensity score will have the same distribution on the set of covariates, here referred to as "covariate balance". Assuming strong ignorability and that covariate balance obtains, estimates of treatment effects from nonrandomized studies approximate those that would have been obtained if randomized studies were conducted instead. In practice, however, the true propensity score is unknown, and instead, propensity scores are estimated by fitting a model such as a logistic regression model of the treatment assignment on the pre-treatment covariates. Then, based on the estimated propensity scores, different techniques, including stratification, weighting, matching and regression adjustment, are utilized in the outcome model to obtain less biased estimate of the causal effect of interest. Another important assumption of propensity score methods is that the potential outcome for each unit remains the same no matter what treatments other units receive. This is also known as the *stable unit treatment value assumption* (SUTVA) (Rubin, 1986). The definition of a useful causal estimand and consistent estimation of the causal estimand may become problematic if SUTVA does not hold.

As noted, there are several common methods for implementing a propensity

score adjustment. The propensity score stratification method (Rosenbaum and Rubin, 1983) incorporates the idea of Cochran (1968) that five subclasses can remove as high as 90 percent of the bias due to the sub-classifying covariate. The stratification method first sorts all the participants by their estimated propensity scores $\hat{e}(X)$ and then stratifies them into multiple mutually exclusive strata (usually five strata) based on the propensity score distribution. Increasing the number of strata should result in improved bias reduction, although the margin of possible bias reduction decreases as the number of strata increases (Cochran, 1968). The treatment effect is first estimated within each stratum and then the overall causal effect is obtained by averaging over all the within-strata treatment effects. To remove the residual imbalance within each stratum, Dehejia and Wahba (1999) proposed a variant of the stratification method that includes the logit propensity scores or polynomials of logit propensity scores into the outcome model within each stratum.

In terms of the variance estimation for the stratification method, a common approach is to simply pool the stratum-specific variance estimates together. However, the uncertainty in the estimated propensity scores is not taken into account by the pooling method. To consider the uncertainty in the propensity score, the sandwich variance estimator can be derived using the M-estimator theory (Lunceford and Davidian, 2004), though its use in practice is limited due to its complexity. The applications of the stratification method has been increasing rapidly in educational and psychological research (e.g., Leow et al., 2004; Yanovitzkya et al., 2005; Swanson et al., 2007).

As another popular propensity score technique, the inverse-propensity weighting approach is based on the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) from the survey sampling literature. The inverse-propensity weighting approach weights participants in the treatment group and the control group by $1/\hat{e}(X)$ and $1/(1-\hat{e}(X))$, respectively, to balance these groups on confounding covariates. This particular weight yields the average treatment effect. To obtain an estimate of

the average treatment effect on the treated, one could utilize the weight

$$T + (1 - T)\frac{\hat{e}(X)}{1 - \hat{e}(X)},$$

where $T = 1$ if the individual is assigned to the treatment group and $T = 0$, if not. Inverse-propensity weighting was first proposed by Rosenbaum (1987) as a form of model-based direct standardization. A variety of other weighting estimators have been developed, including the weighted estimator with standardized weight, the minimum large sample variance weighted estimator and the doubly-robust estimator. The doubly-robust estimator was found to provide an unbiased treatment effect estimate as long as either the outcome model or the propensity score model is correct. The details of these weighting approaches and corresponding variance estimates can be found in Lunceford and Davidian (2004).

Another widely used propensity score method is propensity score matching that matches participants in the treatment group to participants in the control group based on their propensity scores. In order to perform propensity score matching, several decisions have to be made. First, one must choose between matching with replacement and without replacement (Rosenbaum, 2002). For matching with replacement, each subject can be used more than once to form matched sets and the variance estimate needs to take into account that the same subject may contribute to multiple matched sets. The variance calculation for matching without replacement is much easier relative to matching with replacement. However, matching without replacement may encounter the difficulty of forming appropriate matched sets when the number of control subjects is smaller than the number of treated subjects.

A second choice one needs to make is between greedy matching and optimal matching. In the greedy matching, a treated subject to be matched is selected at random and then a matched pair is formed by matching this subject with its nearest neighbor in the control group, that is, the subject in the treatment group is matched with the subject in the control group with the closest propensity scores. The greedy matching may yield overall poor balanced pairs of subjects, which can be overcome by the optimal matching method (Rosenbaum, 1989). The idea of optimal matching

is to search for the optimal construction of matched sets that minimizes the overall distances between the subjects in the matched sets by the way of a minimum-cost network flow solver.

Lastly, one has to decide whether 1:1 matching, 1:M matching, M:1 matching or full matching is desirable. In 1:1 matching, one treated subject is matched with only one control subject. In 1:M (M:1) matching, one treated (control) subject can be matched with M control (treated) subjects. In full matching, one treated subject can be matched with one or multiple control subjects while one control subject can also be matched with one or more treated subjects. Assuming all the cases are included for the analysis, the matching with fixed ratio has the advantage of avoiding severely unbalanced dispersion of treatment units and control units in the matched sets, which leads to more efficient estimators. Pair matching would be the best in this sense as treatment units and controls units are equally balanced, one treatment unit matched to one control unit. But fixed ratio matching has a disadvantage of increasing bias because the total distance between matches is not minimized. The full matching method enjoys the benefit of forming matched sets with the minimum total distance, which yields less biased estimator, but has the potential drawback of forming severely unbalanced dispersion of treatment units and control units in the matched sets and increasing the variance. For example, in order to minimize the total distance among matched sets, full matching may yield a match of one treatment unit to many control units in one match and one control unit to many treatment units in another match. The problem of choosing between fixed ratio matching and full matching essentially is a variance and bias trade-off problem (Hansen and Klopfer, 2006). The applications of the matching approach can be found in Foster (2003), Dehejia and Wahba (2002) and Guo et al. (2006).

Finally, the regression adjustment approach directly includes the propensity score or the logit of propensity score linearly (and possibly non-linearly) in the outcome equation (Kang and Schafer, 2007; Schafer and Kang, 2008). The regression adjustment approach is easy to implement but relies heavily on the correct specification of the regression model. Applications of this approach can be found in Kurth et al. (2006) and Shadish et al. (2008).

1.4 Bayesian Propensity Score Analysis

Over the past two decades, significant advances have been made in the area of Bayesian statistical inference, owing mostly to computational developments and readily available software (e.g., Gilks and Spiegelhalter, 1996). These computational advances have led to increased applications of Bayesian methods to research in the social and behavioral sciences. As is well known, the Bayesian perspective begins by specifying a model for an outcome of interest, elicits prior distributions for all model parameters and obtains the joint posterior distribution of the model parameters given the data via Bayes' theorem and some chosen MCMC algorithm, such as the Metropolis-Hastings algorithm or the Gibbs sampler. For a general review specific to the social and behavioral sciences, see Kaplan (2014).

1.4.1 The Joint-Modeling Bayesian Propensity Score Method

Rubin (1985) argued that a Bayesian approach to propensity score analysis should be of great interest to the applied Bayesian analyst, and yet propensity score estimation within the Bayesian framework was not addressed until relatively recently. Hoshino (2008) argued that propensity score analysis has focused mostly on estimating the marginal treatment effect and that more complex methods are needed to handle more realistic problems. In response, Hoshino (2008) developed a quasi-Bayesian estimation method for general parametric models, such as latent variable models, and developed a Markov chain Monte Carlo (MCMC) algorithm to estimate the propensity score. McCandless et al. (2009) first provided a practical Bayesian approach to propensity score stratification, estimating the propensity score and the treatment effect and sampling from the joint posterior distribution of model parameters via an MCMC algorithm. The marginal posterior probability of the treatment effect can then be obtained based on the joint posterior distribution. Using a simulation study and a case study, McCandless et al. (2009) found that weak associations between the covariates and the treatment led to greater uncertainty in the propensity score and that the Bayesian subclassification approach yields wider credible intervals compared to the frequentist counterpart. Similar to the

McCandless et al. (2009)'s study, An (2010) presented a Bayesian approach that jointly models both the propensity score equation and outcome equation at the same time and extended this one-step Bayesian approach to propensity score regression and single nearest neighbor matching methods.

A consequence of the Bayesian joint modeling procedure utilized by McCandless et al. (2009) and An (2010) is that the outcome variable that is observed after treatment assignment contributes information to the propensity score estimation, which may result in biased causal effect estimates (Zigler et al., 2013). This is especially problematic if the relationship between the outcome and the propensity score is misspecified (McCandless et al., 2010). Also, joint modeling of the propensity score model and outcome model loses an important benefit of the propensity score approach, that is, checking covariate balance before data collection to improve study design. To solve this problem, McCandless et al. (2010) utilized an approximate Bayesian technique introduced by Lunn et al. (2009) for preventing undesirable feedback between propensity score model and outcome model components. Specifically, McCandless et al. (2010) included the posterior distribution of the propensity score parameters as covariate input in the outcome model so that the flow of information between the propensity score and the outcome is restricted. This so-called *sequential Bayesian propensity score analysis* yields treatment effect estimates that are comparable to estimates obtained from frequentist propensity score analysis. Nevertheless, as McCandless et al. (2010) point out, their method is only approximately Bayesian and also encounters the difficulty that the Markov chain is not guaranteed to converge.

1.4.2 The Two-Step Bayesian Propensity Score Approach

In order to maintain a fully Bayesian specification while overcoming the conceptual and practical difficulties of the joint modeling methods of McCandless et al. (2010) and An (2010), a two-step Bayesian propensity score approach was recently developed by Kaplan and Chen (2012) that can incorporate prior information on the model parameters of both the propensity score equation and outcome model

equation. Consistent with Bayesian theory (see e.g., De Finetti, 1974), specifying prior distributions on the model parameters is a natural way to quantify uncertainty – here in both the propensity score and outcome equations. An unpublished manuscript (Chib and Greenberg, 2010) also described the idea of modeling propensity score and outcome equations separately for Bayesian propensity score matching. However, it does not provide details in regard to analytic procedures, especially unclear about the uncertainty estimates. In contrast, Kaplan and Chen (2012) delineated the two-step Bayesian propensity score approach to stratification, weighting and optimal matching through comprehensive simulation studies and a case study and derived the treatment effect and variance estimators. Thus, in this dissertation, I review the two-step Bayesian propensity score approach mainly based on the Kaplan and Chen (2012) paper.

In the Kaplan and Chen (2012) two-step Bayesian propensity score approach (hereafter, BPSA), the propensity score model is fit in the first step, specified as the following logit model.

$$\text{Log}\left(\frac{e(X_i)}{1-e(X_i)}\right) = \alpha + \beta' X_i, \tag{1.4}$$

where α_i is the intercept, β_i refers to the slope and X_i represents a design matrix of chosen covariates for individual i. After the propensity score estimates are obtained, a Bayesian outcome model is fit in the second step to estimate the treatment effect via various propensity score methods such as stratification, weighting, matching and regression adjustment.

If the posterior sampling of a chosen Bayesian logit model has 1000 iterations with a thinning interval of 1, then for each observation, there are $M = 1000$ posterior draws of the propensity score, and based on each posterior draw of the propensity score, $e(X_m)$, there are $L = 1000$ posterior draws of the treatment effect generated from the posterior distribution of γ ($m = 1, \ldots, M$, $l = 1, \ldots, L$), where γ is the treatment effect. Kaplan and Chen (2012) then provides the treatment effect

estimator as follows,

$$E(\gamma \mid X, Y, T) = M^{-1}L^{-1} \sum_{m=1}^{M} \sum_{l=1}^{L} \gamma_l(\eta_m), \tag{1.5}$$

where $L^{-1} \sum_{l=1}^{L} \gamma_l(\eta_m)$ is the posterior sample mean of γ in the Bayesian outcome model based on the m^{th} set of propensity scores η_m and then this posterior sample mean is averaged over M sets of propensity scores. The posterior variance of γ is based on the total variance formula as follows,

$$Var(\gamma \mid X, Y, T) = M^{-1} \sum_{m=1}^{M} \sigma^2_{\gamma(\eta_m)} + (M-1)^{-1} \sum_{m=1}^{M} \{\mu_{\gamma(\eta_m)} - M^{-1} \sum_{m=1}^{M} \mu_{\gamma(\eta_m)}\}^2, \tag{1.6}$$

where

$$\sigma^2_{\gamma(\eta_m)} = (L-1)^{-1} \sum_{l=1}^{L} [\{\gamma_l(\eta_m) - L^{-1} \sum_{l=1}^{L} \gamma_l(\eta_m)\}]^2, \tag{1.7}$$

is the posterior sample variance of γ in the Bayesian outcome model under the m^{th} set of propensity scores and

$$\mu_{\gamma(\eta_m)} = J^{-1} \sum_{l=1}^{L} \gamma_l(\eta_m), \tag{1.8}$$

is the posterior sample mean of γ in the same Bayesian outcome model. The equation (1.6) captures two sources of variation. The first source of variation is the average of the posterior variances of γ across the posterior draws of the propensity score, represented by the first part of the right hand side of equation (1.6), and the second source of variation comes from the variance of the posterior means of γ across the posterior draws of the propensity score, obtained by the second part of the right of hand side of equation (1.6).

Utilizing the above estimators, Kaplan and Chen (2012) conducted three comprehensive simulation studies as well as a small case study comparing frequentist

propensity score analysis with the two-step Bayesian alternative focusing on the treatment effect and variance estimates. The effects of different sample sizes, true treatment effects and choices of priors on treatment effect and variance estimates were evaluated. Consistent with Bayesian theory, Kaplan and Chen (2012)'s findings showed that lower prior precision of the treatment effect is desirable when no prior information is available in order to obtain the estimates similar to frequentist results but with more accurate intervals; and higher prior precision is preferable when accurate prior information regarding treatment effects is attainable in order to obtain more precise treatment effect estimates. For the case of small sample size, the Bayesian approach shows slight superiority in the estimation of the treatment effect compared to the frequentist counterpart.

Kaplan and Chen (2012) studied the treatment effect and variance estimates of the two-step Bayesian propensity score approach, but they did not examine the performance of their approach with respect to covariate balance. A further study of the covariate balance properties of the Kaplan and Chen (2012) approach was given in Chen and Kaplan (2014). Their results of a case study revealed that both Bayesian and frequentist propensity score approaches substantially reduced initial imbalance as expected, and their performance on covariate balance was similar in regard to the standardized mean/proportion differences and variance ratios in the treatment group and control group. Similar performance was also found with respect to the 95% bootstrap intervals and 95% posterior probability intervals. Specifically, although the frequentist propensity score approach provided slightly better covariate balance for the propensity score stratification and weighting methods, the two-step Bayesian approach offered slightly better covariate balance under the optimal full matching method. Results of Chen and Kaplan (2014)'s simulation study indicated similar findings. In addition, the Bayesian propensity score approach with informative priors showed equivalent balance performance compared to the Bayesian approach with noninformative priors, indicating that the specification of the prior distribution in the propensity score model did not greatly influence the balance properties of the two-step Bayesian approach. The optimal full matching method, on average, offered the best covariate balance compared to the

stratification and weighting methods for both Bayesian and frequentist propensity score approaches. Chen and Kaplan (2014) also found that the two-step Bayesian approach under optimal full matching with reliable priors provided, on average, the smallest standardized mean/proportion difference and variance ratio of the covariates between the treatment and control groups.

Chen and Kaplan (2014) argued that a benefit of conducting Bayesian propensity score analysis is that one can obtain a distribution of posterior propensity scores and thus a distribution of corresponding balance indices (e.g. Cohen's d and variance ratio) so that the variation in balance indices can be studied in addition to the point estimates to assist in balance checking. Good balance is achieved if both the point estimates and the posterior probability intervals of the balance indices fall into the desirable range.

The Bayesian propensity score approaches described in the preceding paragraphs all consider samples with independent observations, where each subject can be viewed as independent from other subjects. However, observations in educational research often have hierarchical structure such as students nested in schools so that one subject's treatment selection and/or outcome variable are very likely dependent from others'. In this case, multilevel modeling techniques could be utilized to account for the within-cluster dependencies among observations and will be discussed in the following chapter.

2.1 Multilevel Modeling

Multilevel data structure is very common in education, psychology, social science, public health, etc., where students are nested in schools, clients are nested in clinics, individual are nested in social groups, patients are nested in hospitals, etc. Other examples of hierarchical structure include but do not limit to cross-national studies with individuals nested in nations, marketing research with stores nested in districts, and qualitative research with respondents nested in interviewers. Some special and less obvious applications of hierarchical models are longitudinal studies, where a series of repeated measures across time are viewed as nested in individuals, and meta-analysis, where the subjects are treated as nested within studies. Through this dissertation, I mainly focus on a two-stage sampling design of multisite studies, where the upper-level clusters are first randomly sampled and then the lower-level subjects are sampled from the available clusters, e.g., students within schools. Nevertheless, the models discussed for the two-level case can be extended to a much larger range of analysis such as higher-order models and growth curve models.

In the multilevel context, the observations within a cluster tend to be more similar compared with those observations in other clusters, which demonstrates dependencies among individuals within clusters. To quantify the dependencies, a useful measure is the intra-class correlation coefficient (ICC), which models the degree of similarity of observations within clusters (Snijders and Bosker, 2012). The ICC can be defined as the ratio of population variance between clusters and the total variance, expressed as $\tau^2/(\tau^2 + \sigma^2)$, where τ^2 represents the between-cluster or cluster-level variation and σ^2 refers to the within-cluster or individual-level variation. The higher the ICC values are, the stronger resemblance the observations have within the same cluster and/or the more apart the clusters are from each other. When the ICC values are not negligible, ignoring the hierarchical structure may lead to serious underestimation of the variance due to the violation of the independence assumption. In fact, the cluster effect is often non-ignorable because

a large cluster size can amplify the small within-cluster dependencies, which results in underestimated variance and produces spuriously "significant" results.

A common procedure of dealing with the hierarchical data is to aggregate the variables at the lower level to the upper level, e.g., using the cluster mean as an aggregated variable, and then analyzing data at the cluster level. The aggregation method is convenient to use, but encounters several problems. Statistically, the sample size is greatly reduced when aggregating data from the individual level to the cluster level and thus a large amount of information is lost. The statistical power is thus weakened. Conceptually, the relationships among the lower-level variables may be different or even opposite to the relationships among the aggregated cluster-level variables, which is known as the *ecological fallacy* or *Robinson effect* (Robinson, 1950). Therefore, results based on the aggregated data can only be interpreted at the upper level.

The other procedure that reconstructs the hierarchical data into a single-level structure is disaggregation, where the cluster-level variables are moving to the individual level and the analysis is done with the individual-level variables. However, the disaggregation method also has some limitations. Statistically, the sample size increases largely after the disaggregation. As the ordinary statistical test treats all the disaggregated data as independent individual-level units, the type I error is seriously inflated. Conceptually, making inference at the upper level using the lower-level variables can be misleading, which is called the *atomistic fallacy*. Therefore, for the hierarchical data, performing analysis on only one of the levels is inappropriate. Instead, all levels should be kept and analyzed in their own ways. Multilevel modeling was born of this, which can properly model structural relationships among variables at different levels and also accurately account for uncertainty in prediction and estimation.

2.2 Specification of Multilevel Models

The pervasiveness of multilevel data has boosted the development of corresponding statistical methods, with the same framework but under different names such

as multilevel models in sociological research, hierarchical models in educational research, and mixed effect models or random effects models in statistical and biostatistical research. Throughout this dissertation, I use multilevel modeling and hierarchical modeling interchangeably in order to reflect the structural feature of the multilevel data.

The general formulation of the multilevel modeling was first laid out by Lindley and Smith (1972), where the unknown parameters in the individual-level model serve as dependent variables in the cluster-level model. I denote the multilevel model with only random intercept as the random intercept model (RI) and the model with both random intercept and slopes as the random intercept and slope model (RIS). For example, assume that there are J schools in the sample, with n_j first-grade students from each school. At the student level, there is one dependent variable, reading score at the end of the first-grade (Y), and two independent variables, how often parents read to children per month (X_1) and students' previous reading scores at the beginning of the first-grade (X_2). At the school level, there is one predictor, percentage of free lunch in each school (W), which reflects the average social economic status of students in each school. Then, the level-1 model can be specified as

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \epsilon_{ij}, \tag{2.1}$$

where Y_{ij}, X_{1ij} and X_{2ij} refer to the reading score, the monthly frequency of parents' reading to children and the previous reading score of the i^{th} student at the j^{th} school, respectively. The coefficient β_{0j} is the intercept and β_{1j} and β_{2j} are the slopes of two predictors, respectively. The coefficient ϵ_{ij} refers to the level-1 residual and is assumed to be normally distributed with mean zero and variance σ^2. The level-1 residual represents the within-cluster variation.

Assuming that only the intercept in the level-1 model varies across different schools with fixed slopes, then this model is denoted as a random intercept model. In the level-2 model as follows, the random coefficients can be modeled using the

school-level predictor W:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{2.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j, \tag{2.3}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} W_j. \tag{2.4}$$

In equations (2.2) to (2.4), γ_{00}, γ_{01} and γ_{00} are fixed intercepts and γ_{01}, γ_{11} and γ_{21} are fixed slopes. The term u_{0j} refers to the variation of the random intercept, which is usually assumed to be normally distributed with mean zero and be independent of the level-1 residual ϵ_{ij}.

In addition to random intercept, if the slopes in equation (2.1) are believed to vary across schools, then a random intercept and slope model can be specified as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{2.5}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, \tag{2.6}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} W_j + u_{2j}. \tag{2.7}$$

The random terms u_{0j}, u_{1j} and u_{2j} are the level-2 residuals, which are often correlated. The variance-covariance matrix of u_{0j}, u_{1j} and u_{2j}, denoted as Δ, reflects the between-cluster variation.

In the hierarchical linear model as the example described above, the outcome variable at the level-1 model is continuous and the level-1 residual is reasonable to be assumed normally distributed. However, for some other types of dependent variables such as dichotomous variable, count data and categorical variables, the assumption of normal residuals is violated and thus the level-1 model needs to be transformed into a different form. Instead of directly modeling the non-normal dependent variables, the expected values of the outcome variables are transformed via a link function (McCullagh and Nelder, 1989). The transformed variable is often assumed to be normally distributed and serves as the new dependent variable in the level-1 model. Take an example of the binary outcome, the logit function is

commonly used as the link function to transform the expected value p into a new variable. If the same predictors were used as in the above example, then the level-1 model would be

$$\mathrm{Logit}(p_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij}. \tag{2.8}$$

Other non-normal level-1 models include the log link for count data, the cumulative logit link for ordinal outcomes and the multinomial logit link for multi-category nominal outcomes. The level-2 models still remain linear as in equations (2.1) to (2.4) because the latent random intercept and slopes are assumed to be normally distributed. This class of hierarchical models with non-normally distributed outcome variables is actually the multilevel extension of the generalized linear models, which is often called multilevel or hierarchical generalized linear models.

2.3 Estimation of Model Parameters

The estimation methods of the fixed effects and variance-covariance components in multilevel modeling include the likelihood-based approaches, i.e., maximum likelihood (ML) and reduced maximum likelihood (REML), and the full Bayesian approach. For estimating the random effects, the empirical Bayes method or full Bayesian approach can be utilized. In this section, I review the likelihood-based methods for estimating model parameters and the empirical Bayes method for estimating the latent random effects, while the full Bayesian approach will be discussed in the next section of the Bayesian hierarchical modeling.

Maximum likelihood is the most commonly used estimation method in multilevel modeling, which estimates the parameters by maximizing the likelihood of the data. Employing the maximum likelihood estimation is beneficial mainly in terms of three aspects. First, it is generally robust to the mild assumption violation with large samples (Hox, 2010). Second, it produces estimates that are asymptotically efficient and consistent, i.e., the maximum likelihood estimates approximate the true parameter with minimum variance when sample size is infinitely large. Third, the sampling distribution of the ML estimates is asymptotically normal, which

facilitates the construction of confidence intervals and statistical tests (Raudenbush and Bryk, 2002). Overall, the ML method is asymptotically optimal. However, a major drawback of the maximum likelihood estimation is that the ML variance estimates are conditional on the estimated fixed effects but the ML method does not take into account the loss of degrees of freedom due to estimating the fixed effects. As a result, the variance tends to be underestimated, which leads to short confidence intervals and overly liberal statistical tests (Raudenbush and Bryk, 2002). This is especially problematic when the number of clusters is small.

The problem caused by ignoring the uncertainty of the fixed effects in the ML estimation can be solved by the restricted maximum likelihood (REML) estimation, which was originally described by Patterson and Thompson (1971). The REML method estimates the variance components after removing the fixed effects from the model (Searle et al., 1992), which accounts for the loss of degrees of freedom due to the fixed effects and provides less biased variance estimates (Longford, 1993). For the two-level model, the REML method generally produces similar level-1 variance estimates to the ML method (Raudenbush and Bryk, 2002), but when the number of upper-units is small, the impact of the uncertainty in the fixed effects becomes stronger and thus the difference between the REML estimates and the ML estimates tends to be greater (Dedrick et al., 2009). Nevertheless, the ML method is still widely used in practice by virtue of the easier computation and more importantly, the ability of comparing the hierarchical models with differences in the fixed coefficients because all the regression coefficients are included in the likelihood function. For the REML method, only differences in the random variance components can be compared (Hox, 2010).

Though the general framework of the hierarchical model was proposed by Lindley and Smith (1972), it received little attention due to the limitation in computing the ML and REML estimates. A breakthrough was made by a seminal paper Dempster et al. (1977), which offered a powerful iterative algorithm consisting of an expectation step followed by a maximization step, so-called the *EM algorithm*. It is broadly applicable and conceptually feasible for the maximum likelihood estimation with the presence of missing data or latent variables. Dempster et al. (1981)

demonstrates the application of the EM algorithm to the parameter estimation in the hierarchical models. The EM procedure has been shown to assure convergence over a broad range of conditions (Wu, 1983). But a big minus of the EM algorithm is that it converges slowly to the ML or REML estimates (Draper, 1995). Nonetheless, combining another iterative algorithm, the Fisher scoring algorithm (Longford, 1987), with the EM algorithm can accelerate the convergence process. The mixture of these two algorithms has been incorporated into the HLM software (Raudenbush et al., 2004). In addition, other iterative algorithms such as the iterative generalized least square (IGLS) (Goldstein, 1986) and the Newton-Raphson algorithm (Lindstrom and Bates, 1988) have been developed and utilized for the parameter estimation in multilevel modeling. All these algorithms can be employed for both the ML and the REML estimations.

In the parameter estimation of multilevel models, the latent random effects are integrated out and can not be obtained through the ML or REML estimation methods discussed above. However, a method known as *empirical Bayes* estimation provides a comprehensive solution to this estimation problem (e.g. Efron and Morris, 1975; Raudenbush and Bryk, 1986). Basically, it combines two kinds of information to estimate the cluster-specific effects: the prior knowledge of the group effect based on the population, e.g., grand mean, and the observed data information from group j, e.g., group mean or the ordinary least square estimate. It is *Bayesian* in terms of utilizing the prior knowledge for estimation and it is *empirical* Bayesian due to the fact that the prior is estimated based on the data, rather than a pre-specified prior probability distribution. One of the benefits of the empirical Bayes approach is that the outliers owing to a large sampling error in observed data can be controlled by the population information. Empirical Bayes shrinks the outlier values toward the population mean and thus offers smaller mean squared error than the ordinary least square estimates Efron and Morris (1975). The empirical Bayes method is available in the software packages such as HLM Raudenbush et al. (2004). More details could be found in Snijders and Bosker (2012).

In terms of the practical implementation, in recent twenty years, multilevel modeling has been incorporated in various softwares and packages such as HLM

(Raudenbush et al., 2004), MIXOR(Hedeker1996), MLwiN (Rasbash et al., 2000), lme4 (Bates and Sarkar, 2007) in R (R Development Core Team, 2012), SAS PROC MIXED (SAS Institute, 2008) and VARCL (Longford, 1990), which makes its use more widespread. However, for more complicated hierarchical models such as nonlinear hierarchical models, the estimation methods in many softwares are still under development. In this case, by taking a distinct perspective from the frequentist school of estimation such as the full or restricted maximum likelihood estimation methods, Bayesian estimation approach seems more appealing, which is reviewed in the following section.

2.4 Bayesian Hierarchical Modeling

Bayesian perspective is natural for the hierarchical modeling, which is evidenced by Lindley and Smith (1972), a Bayesian article that first laid out the framework of hierarchical models. In the conventional multilevel modeling, the unknown parameters characterize a population, from which the data of interest are sampled with certain probability. All the parameters are viewed as fixed and are estimated based on the sampled data. Only the upper-level effects are treated as latent random variables. In contrast, from a Bayesian perspective, all statistical parameters are viewed as random and have probability distributions of their own that quantify an investigator's uncertainty about the parameter values. The parameters of interest at different hierarchical levels can be modeled and estimated through updating a prior distribution that reflects a researcher's belief using the observed data. If little information is available before collecting the data, a diffuse prior can be utilized. However, in many cases, at least the parametric family of the prior distribution can be assumed and the unknown hyperparameters such as mean and variance in a normal prior distribution can be specified by hyperpriors (Gelman et al., 2003). The hierarchical structure of prior and hyperprior demonstrates the hierarchical nature of the Bayesian models.

Though Bayesian hierarchical models represent a much larger class of analytic models, in this section, Bayesian hierarchical modeling particularly refers to the

application of Bayesian models into hierarchical data. For a Bayesian hierarchical model in the multilevel context, the random effects are random parameters with their own prior distributions. The unknown parameters in the prior distributions, i.e., hyperparameters, are then specified by hyperpriors, that is, the prior information for the hyperparameters. Specifically, using the notations in equation (2.2), the random intercept β_{0j} is treated as a parameter that has a prior distribution with the prior mean as a function of the upper-level covariates and the prior variance as the variance of u_{0j}. Then the unknown hyperparameters in the prior distributions, i.e., γ_{00},γ_{01} and u_{0j} in equation (2.2), are assigned hyperprior distributions. The random slopes β_{1j} and β_{2j} can be specified in similar ways.

Generally, Bayesian techniques provide the following unique advantages for estimating parameters in multilevel models: (1) From a Bayesian perspective, reliable prior information can assist with parameter estimation, which is especially desirable in the case of small sample size. The prior distribution can come from results based on prior studies (e.g. prior circle of studies and meta-analytic studies) or expert opinions before any data is collected; (2) When it is inappropriate to assume higher-level units are randomly sampled from some population (e.g., all the European countries), Bayesian statistics avoids the philosophical problem because it is the parameter, not the data, that are sampled (Sutton, 2012); (3) Bayesian approaches naturally and fully account for uncertainty for all the parameters and provide any information needed for inference once the posterior sample is obtained, while the variance estimation via likelihood-based methods can sometimes be very complex and only provide asymptotic solution with unfeasibly large sample size. However, as a trade-off, sampling from the posterior distributions is usually much more time-consuming than the likelihood-based methods. But in the case of complicated hierarchical models, the difference in computing time between Bayesian method and likelihood-based method shrinks. Another flip side of Bayesian approaches is that the prior and hyperprior distributions needs to be chosen carefully such as specifying priors based on previous studies or expert opinions. If there is not much reliable prior information, a diffuse prior would be preferable. Incorrect priors with high prior precisions can produce seriously distorted estimates (Kaplan and Chen,

2012).

In light of the advantages of Bayesian methods, this dissertation explores the use of Bayesian hierarchical modeling, in particular, Bayesian hierarchical logit model, in multilevel observational studies via propensity score analysis. The existing research in propensity score analysis for multilevel observational studies are discussed in the following section 2.5 and then the Bayesian multilevel propensity score analysis is outlined in Chapter 3.

2.5 Propensity Score Analysis in the Multilevel Context

2.5.1 Overview

Nonrandomized experiments or observational studies in educational and medical research often have hierarchical structure. For example, students who self-select into an educational program are almost always nested in schools. Similarly, patients who volunteered to participate in a medical research are often nested in physicians or hospitals. In terms of examining the causal effect in observational studies, propensity score method is most commonly used to reduce selection bias in estimated treatment effects. However, the nonrandomized trials in the multilevel context raise extra methodological challenge because an individual's treatment selection may influence the treatment choices and potential outcomes of other individuals in the same cluster such that the assumption of stable unit treatment value is violated. In addition, not only potential outcomes are dependent for individuals within the same cluster, but also the selection mechanisms may reply on cluster characteristics and vary substantially across clusters.

Take the causal effect of students' grade-retention on their later academic achievement as an example. In addition to the problem of selection bias as mentioned at the beginning of the dissertation, students are almost always nested in schools, which often invokes dependencies among students within the same school. For instance,

a disruptive student may affect other students's learning outcome. Also, students in the same school share similar resources and are managed by the same policy, which can vary greatly from another school. As a result, the academic performance of students within the same school tend to be more correlated with each other than with the achievement of students from other schools and also the retention policy of one school may distinguish from other schools'. The multilevel structure on both outcome and treatment selection raise the need of combining propensity score approach and multilevel modeling to produce less biased treatment effect estimates and more accurate variation estimates.

Though there is a quite large amount of nonrandomzied nested trials, only a very limited number of studies explored the application of multilevel modeling into propensity score analysis to account for the hierarchical structure of treatment selection and potential outcome to reduce selection bias as well as balance confounding covariates. To my best knowledge, in addition to three dissertation work (Kelcey, 2009; Lingle, 2009; Rickles, 2012) and one editorial (Griswold et al., 2010), there are six methodological manuscripts (Hong and Raudenbush, 2006; Arpino and Mealli, 2011; Thoemmes and West, 2011; Kim and Seltzer, 2007; Li et al., 2013; Steiner et al., 2013) with regard to multilevel propensity score approaches in the nonrandomized multisite studies. I first briefly review these six articles in *section 2.5.2* and then discuss treatment selection mechanism, model specification, matching type and study method utilized in these papers in *section 2.5.3*. Specifically, with respect to the treatment selection, two types of mechanisms are examined, treatment selection occurred at only the individual level or treatment selection taken place at both the individual level and cluster level. The model specification part starts with the description of the multilevel propensity score equation and then elaborates different methods for multilevel observational studies discussed by the literature. In terms of the matching type, the performance of different analytic methods for nonrandomized nested designs are investigated in regard to matching within clusters and matching across clusters through evaluating covariate balance and/or accuracy and precision of treatment effect estimation. Last but not the least, the above literature is reviewed according to the study method, namely, simulation

study, case study and both. Other study methods such as within-study comparison and study review have not yet been found in the multilevel propensity score literature.

2.5.2 Existing Research in Multilevel Propensity Score Approach

Hong and Raudenbush (2005) first utilized a multilevel propensity score stratification approach to evaluate the effects of the kindergarten retention policy using the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K) data and the proposed approach was presented in detail in Hong and Raudenbush (2006). In the proposed multilevel propensity score stratification approach, a two-level logistic regression model was first utilized to estimate students' propensity scores of being retained in kindergarten. Then students were stratified based on the estimated propensity scores and the stratum indicators were included in a two-level outcome model with random intercepts and slopes to account for dependencies among the outcomes within schools while balancing student-level and school-level covariates. The kindergarten retention effects were examined for high-retention schools and low-retention schools separately but using similar propensity score and outcome models. Results indicated that propensity score adjusted retention effect estimates were much smaller than the raw retention effect without any adjustment, implying that the multilevel propensity score stratification method effectively reduced selection bias. In addition, Hong and Raudenbush (2006) compared the effects of being in high-retention schools relative to being in low-retention schools on children's reading and math achievement. A single-level logistic regression model was specified to estimate schools' propensity scores of adopting a high retention rate. Then a two-level random intercept model was employed to evaluate the impact of school retention policy on students' achievement controlling for school-level propensity scores.

Covariate balance was checked by comparing the mean and standard deviation of the logit of student-level propensity scores between promoted children and retained children and the logit of school-level propensity scores between high-

retention schools and low-retention schools. After propensity score adjustment, no significant differences were found in the logit propensity scores between different treatment conditions and within-stratum balance is achieved for most of the covariates. In Hong and Raudenbush (2006), a sensitivity analysis was also conducted, which revealed that an unmeasured confounder, at the individual level or school level, comparable to the strongest measured confounder would be sufficient for altering the conclusion about the retention effect. This implies the importance of holding the strong ignorability assumption in multilevel propensity score analysis.

Overall, Hong and Raudenbush (2005, 2006) are the pioneers in regard to extending the propensity score analysis to the multilevel settings and provide a general framework for making causal inference in multilevel observational studies. Results indicated that the proposed multilevel propensity score stratification method greatly reduced the imbalance due to treatment selection, but was sensitive to strong unobserved confounders.

Arpino and Mealli (2011) also examined the use of propensity score matching techniques in multilevel observational studies with the presence of unobserved cluster-level covariates, but focused on estimating the average treatment effect on the treated (ATT). Four types of propensity score models were compared through comprehensive simulation studies, including single-level logit model with covariates at both levels, single-level model with level-1 covariates only, fixed effect logit model with school indicators/dummies and multilevel random intercept model. The robustness of different propensity score models to a omitted cluster-level covariate was evaluated through their performance on the treatment effect estimation and covariate balance. The findings revealed that the omitted cluster-level variable have the strongest influence when it is highly correlated with the potential outcome. When the omitted variables only affect the treatment selection but have no influence on the potential outcomes, controlling for cluster effects is not necessary. Both random effect and fixed effect models capture the unobserved heterogeneity due to omitting the cluster-level covariates quite well. In particular, the propensity score matching using the fixed effect model with school dummies performs the best under all different simulation conditions, which might be limited to the specific

simulation settings. Overall, Arpino and Mealli (2011) demonstrated the influence of omitting cluster-level variables and provided guidance for modeling multilevel observation data sets when there are unobserved covariates.

Similar to Arpino and Mealli (2011), Thoemmes and West (2011) studied four types of propensity score models for reducing selection bias in the multilevel observational studies, but more comprehensively through a simulation study and a case study and also discussing the appropriate use of each model. The propensity score models investigated by Thoemmes and West (2011) include a single-level propensity score model with cluster-level and individual-level covariates as well as across-level interactions between the covariates, a fixed effect model with school dummy indicators, a standard multilevel model that accounts for between-cluster variations in treatment assignment and outcomes and a multilevel model on the *broad inference space*, where clustering is just an incidental feature and treatment selection does not vary across clusters.

The simulation study varied the estimation strategy, matching type, overall sample size and the degree of intra-class correlation (ICC), totally 32 simulation conditions. Results revealed that in the low ICC scenario, the performance of different propensity score models and choices of matching types were subtle, while in the high ICC condition, the regular multilevel propensity score model using both cluster and individual covariates slightly outperformed other models in terms of bias, mean square error and coverage rate under all kinds of sample sizes. Overall, the multilevel propensity score model using the within-cluster matching strategy provided the least biased estimate. The fixed effect model with school indicators together with within-cluster matching also performed very well on bias reduction and balance. However, the within-cluster matching could fail if there were extreme imbalances between treated and untreated individuals such that many subjects could not find suitable matches within clusters. This happened in the case study, where the fixed effect model with school indicators was not estimable. In this case, across-cluster matching would be a desirable alternative because individuals can be matched across different clusters to achieve overall balance. Another finding of the simulation study is that all the propensity score models provided the coverage

rates lower than the 95% nominal level across all simulation conditions, indicating that the variation in treatment effect estimates was underestimated, which may be because the uncertainty of propensity score estimates was ignored in the variance estimation of the treatment effect.

Kim and Seltzer (2007) also investigated the use of different propensity score models in the multilevel settings, including single-level model, random intercept model with fixed slopes and random intercept and slope model, using the technique of within-cluster matching. The above three models were compared through a case study of the Early Academic Outreach Program (EAOP) designed to support academic enrichment and informational access for students who have potential for higher education. Nearest neighbor matching with caliper was adopted to match students in the treated group and the control group within the same school.

Results showed that omitting level-2 covariates in the random intercept and slope model does not have a substantial effect on the matching results. Matching with both random intercept and random slopes yield the best balance among the three models. However, these conclusions are limited to the real data used in the case study and more comprehensive simulation studies are needed to confirm the findings. In fact, through simulation studies, Arpino and Mealli (2011) showed that omitted cluster-level variables that are highly correlated with the potential outcome can lead to largely biased estimates. Consistent with Thoemmes and West (2011), Kim and Seltzer (2007) also described the limitation of the within-cluster matching method, which often fails in the case of small sample sizes. Across-cluster matching can be an alternative when the within-cluster matching is unfeasible.

Interestingly, Kim and Seltzer (2007) also briefly demonstrated a Bayesian treatment of the multilevel random intercept and random slope model and provided the posterior distribution of the propensity scores and the treatment effect using the Markov-Chain Monte-Carlo (MCMC) sampling technique. However, the uncertainty in the propensity score estimates was not taken into account in the treatment effect estimates since only the marginal posterior mean propensity scores are utilized in the outcome model, which underestimated the uncertainty in treatment effect estimates and may result in spuriously significant results.

Li et al. (2013) investigated five types of propensity score weighted estimators for multilevel observational data, that is, marginal estimator without consideration of the cluster structure, cluster-weighted estimator, and three doubly-robust estimators with single-level outcome model, fixed-effect outcome model and random-effect outcome model, respectively. Three kinds of propensity score models were examined, such as single-level model with level-1 and level-2 covariates, fixed effect model with cluster indicators and level-1 covariates, and random intercept model with level-1 and level-2 covariates.

The performance of the combination between the three propensity score models and the five propensity score weighted estimators was examined through simulations under the setting of a random intercept propensity score model and a random intercept and slope outcome model. The authors provided several findings, two of which are summarized here. First, the fixed effect and random effect propensity score models yielded comparable bias in treatment effect estimates given the same outcome model. However, the random effect outcome model outperformed the fixed effect model in treatment effect estimation if coupled with the same propensity score model. Second, for doubly-robust estimators, the correct outcome model with a misspecified propensity score model provided less biased treatment estimates than the correct propensity score model but with a misspecified outcome model. Based on these two findings, the authors draw a conclusion that the outcome model had much larger influence on the final treatment effect estimates than the propensity score model. However, this conclusion is also based on the specific simulation setting, where data were generated from an outcome model that was more complicated than the propensity score model. If the generating propensity score model also has both random intercept and random slopes, the above conclusion might not hold. Also, Li et al. (2013) did not show the uncertainty estimates when comparing the performance of different methods in addition to the treatment effect estimates. Uncertainty estimate is central to statistical inference and it is of interest to confirm that random effect models can better capture the uncertainty than fixed effect models.

Focusing on different matching strategies, Steiner et al. (2013) evaluated and

discussed the use of within-cluster and across-cluster matching to tangle with various practical challenges in multilevel observational studies. Through a simulation study where treatment selection and outcome vary across clusters, Steiner et al. (2013) showed that when propensity score model was correctly specified, that is, consistent with the generating model, within- and across-cluster matching provided comparable treatment effect estimates. When propensity score model was misspecified, within-cluster matching still recovered the treatment effect well, but across-cluster matching produced severely biased treatment effect estimates. These results indicated the advantage of within-cluster matching over across-cluster matching, as within-cluster matching effectively balances all the cluster-level covariates. However, within-cluster matching tends to fail in practice due to lack of overlap in treatment and control units as well as reduction in sample size while matching within each cluster. Steiner et al. (2013) found that across-cluster matching could be but not always employed to compensate the lack of overlap in treatment and control groups. In general, if propensity score model is correctly specified and there are sufficient overlap within each cluster, across-cluster matching can yield consistent treatment effect estimates.

2.5.3 Issues in Propensity Score Analysis for Multilevel Observational Studies

2.5.3.1 Treatment Selection

In multilevel observational studies, treatment selection can occur at the individual level, cluster level, or both levels. Among the literature reviewed above, most studies only investigated the individual-level treatment selection with an exemption of Hong and Raudenbush (2006), where the treatment selection were at both levels. For the individual-level selection that varies across clusters, multilevel propensity score model provides the best balance and treatment effect estimates (Thoemmes and West, 2011; Steiner et al., 2013). Specifically, Kim and Seltzer (2007) described two common multilevel treatment selection mechanisms: (1) random intercept

selection process, which views the effects of individual characteristics on the probability of treatment assignment as fixed and the between-cluster variation solely comes from the school membership; and (2) random intercept and slope mechanism, which views the cluster-level variation coming from the interaction between the cluster membership and individual characteristics in addition to the variation from the cluster membership. For the random intercept selection mechanism, fixed effect model with school indicators performs equally well as the random intercept model, but for the random intercept and slope settings, random slopes have to be included in the multilevel propensity score model in order to avoid misleading estimates (Kim and Seltzer, 2007). For the cluster-level treatment selection, new challenges emerge. For example, the number of clusters is often small relative to the total sample size, which could lead to poorly estimated cluster-level propensity scores. Also, unobserved cluster-level covariates can not be controlled. Two restrictions can be imposed in order to facilitate the causal inference at the cluster level such as no interference among clusters and strong ignorability assumption at the cluster level (Hong and Raudenbush, 2006). For the treatment selection occurring at both levels, the causal inference becomes more complicated. There may be cross-level interactions between cluster-level treatment assignment and subject characteristics or subject-level treatment assignment. Hong and Raudenbush (2006) considered the individual-level treatment effects for high-retention schools and low-retention schools separately, which provided a way to handle with the inference in this case. Further research on analyzing all clusters together but still accounting for treatment selection at both levels would be beneficial.

2.5.3.2 Model Specification

Propensity score analysis is a two-stage process where propensity score is estimated at the first stage and then outcome model is specified at the second stage based on the propensity score adjustment. For the multilevel observational studies, the propensity score models discussed in the preceding literature mainly include single-level logit models with covariates, fixed effect logit models with school

dummy variables and standard multilevel logit models with random intercept or with both random intercept and slopes. Since treatment selection is usually affected by covariates at both individual level and cluster level, the propensity score in the multilevel context can be generally expressed as $e(X_{ij}) = P(T_{ij} = 1 \mid X_{ij}, W_j)$, which represents the probability of being in the treated group for individual i at cluster j given a vector of observed individual-level covariates X_{ij} and cluster-level covariates W_j.

According to Griswold et al. (2010), the above models for estimating propensity scores in multilevel observational studies can be classified into three categories: (1) Ignore potential clustering completely, that is, complete pooling of covariate information across clusters by using a single-level logistic regression model without school indicators. The assumption underlying this approach is that individuals in different clusters with similar characteristics would have similar probabilities of receiving treatment and thus all individuals' information are used for the propensity score estimation. In this case, the covariate balance after propensity score adjustment is often checked across clusters and not within each cluster because the cluster effect is totally ignored and the propensity score estimates may be biased within single clusters (Thoemmes and Kim, 2011). (2) Model the propensity score completely within clusters, that is, no pooling across clusters. A separate logistic regression model is fitted for each cluster, where subjects are compared within clusters and thus the unbalance in observed and unobserved cluster-level covariates is removed. Balance is often checked within clusters and can also be checked across clusters by combining the balance indices of matched subjects within each cluster. However, a major drawback of this approach is the lost of statistical power since the propensity score estimation is based on each cluster alone. (3) Use multilevel logistic models to include cluster-specific information, that is, partial pooling information across different clusters. Subjects need not to be in the same cluster to be compared, but can be from different clusters with similar cluster-level covariates. Such approaches usually achieve better within-cluster balance relative to overall balance as it approximates multisite randomized trials where individuals are randomized within each cluster (Thoemmes and Kim, 2011).

After the propensity score adjustment through stratification, weighting or matching techniques, an outcome model can be specified as fixed effect model or multilevel model. Li et al. (2013) showed that utilizing a multilevel outcome model in multilevel observational studies can help with reducing bias in the treatment effect estimates. One limitation for all the reviewed literature is that the uncertainty estimate of the treatment effect was not laid out clearly and the uncertainty of propensity scores is very likely to be ignored in the variance estimation.

2.5.3.3 Matching Type

In multilevel settings, propensity score can be employed through both within-cluster matching and across-cluster matching. For within-cluster matching, subjects are only matched within each cluster based on their propensity scores and are not allowed to match with subjects from other clusters. Within-cluster matching using the propensity score intends to approximate blocked randomized design, where, for example, schools are first randomly sampled from a population, and then within each school, students are randomly selected and assigned to different treatment conditions. In multilevel observational studies, students self-select into the treatment group and control group and very likely have imbalanced covariate information. Through within-cluster matching, students are matched on propensity score within each school to recover the blocked randomized trials where students' background covariates are balanced in different treatment conditions in each school.

In contrast to within-cluster matching, across-cluster matching matches subjects regardless of their cluster memberships. Subjects in one cluster can be matched with subjects in the same cluster as well as subjects in other clusters so that information from other clusters can be borrowed. A good match using the across-matching method requires all cluster-level covariates that are related with the treatment selection and the outcome are included in the propensity score model. Failing to account for important covariates at the cluster level can yield very poor matches when matching across clusters and in turn produce severely biased treatment effect estimates.

Thoemmes and West (2011) pointed out that propensity score models may perform distinctly for different matching types. Particularly, single-level propensity score models with covariates are expected to perform better for across-cluster matching and would achieve good overall balance if treatment selection does not vary across clusters. Fixed effect models with school dummy indicators and multilevel propensity score models approximate multisite randomized trials and are suitable for within-cluster matching. These two kinds of models are similar in terms of allowing the treatment selection varying across clusters. However, Steiner et al. (2013) showed that given correctly specified propensity score model, both within- and across-cluster matching can yield consistent treatment effect estimates.

2.5.3.4 Study Method

In a simulation study, data is generated with known properties and various conditions can be examined based on the purposes of research studies. A great benefit of a simulation study is that the true parameter value such as treatment effect is known so that the performance of different models or methods can be compared with the absolute benchmark. As a trade-off, a simulation study may have settings that are far from the real practice and may not guide the practitioners directly. In contrast, a case study compares methods on one or more real data sets. In contrast to simulation studies, the true treatment effect in a case study is unknown and thus different methods can only be compared relatively. An advantage of case studies is that data are real and thus can straightly guide the practice.

In terms of the propensity score analysis in the multilevel settings, Hong and Raudenbush (2006) and Kim and Seltzer (2007) only conducted case studies to demonstrate the proposed multilevel propensity score method and compared different propensity score models in the multilevel context, respectively. Arpino and Mealli (2011) utilized simulation studies alone to examine the effects of omitted cluster-level variables on treatment effect estimation and covariate balance through different propensity score models. Li et al. (2013) and Thoemmes and West (2011) adopted both simulation studies and case studies to illustrate and compare different

propensity score models in multilevel observational studies for weighting and matching, respectively. Through the combination of simulation studies and case studies, one is able to identify the models that perform the best in terms of bias and balance, such as the multilevel propensity score models (e.g., Steiner et al., 2013; Li et al., 2013; Kim and Seltzer, 2007), and also are able to realize the practical limitation of certain models, such as the fixed effect model with school indicators, for small data sets (Thoemmes and West, 2011).

The methodological literature reviewed above focused on the analytic use of multilevel propensity score methods. Few literature is available from a design perspective to examine the use of multilevel propensity score methods. Also, most literature of multilevel propensity score analysis are limited to dichotomous treatment conditions. More general treatment conditions such as multiple treatments or continuous treatments can be further studied. Last but not the least, the uncertainty of propensity score estimates is often ignored in the multilevel propensity score literature. A Bayesian approach to multilevel propensity score analysis could be a solution to this problem, which is elaborated in Chapter 3 below.

3 TWO-STEP BAYESIAN MULTILEVEL PROPENSITY SCORE APPROACH

3.1 Method

A review of the extant literature shows that Bayesian propensity score analysis has not been extended to the multilevel settings. A unique benefit of Bayesian multilevel propensity score approach is the ability of naturally accounting for uncertainty. Due to the dependencies within clusters in the multilevel data and the two-step nature of the propensity score procedure, uncertainty in propensity score estimation in the multilevel context is hard to be taken into account within the frequentist framework and is often ignored. This dissertation extends the two-step Bayesian propensity score approach proposed by Kaplan and Chen (2012) into the multilevel settings and provides a practical Bayesian propensity score approach for making causal inference in multilevel observational studies while accounting for uncertainty in both the propensity score and the outcome. In addition, through examining different sample sizes, intra-class correlations, priors, matching strategies, model specifications and propensity score methods, this dissertation provides researchers some guidance on selecting an appropriate matching technique.

The proposed Bayesian multilevel propensity score approach adopts a two-step procedure. In the first step, the posterior propensity scores are obtained via a Bayesian multilevel logit model with either random intercept only or both random intercept and random slopes, as shown in equations (2.2) to (2.4) and (2.5) to (2.7), respectively. Specifically, the program *MCMChlogit* in the R package *MCMCpack* (Martin et al., 2011) was employed to obtain the posterior samples of the propensity score model parameters for simulation studies. The R package *rjags* (Plummer, 2014) were utilized for the case study, which ran *JAGS* (Plummer, 2003) in R and was found to provide more stationary posterior samples on the real data used in this dissertation. Note that in the frequentist multilevel modeling framework, parameters are either fixed or random. However, the Bayesian framework clarifies the language

in that all parameters are unknown and hence are considered random and assigned probability distributions to reflect the analysts' uncertainty. Nonetheless, this dissertation still uses the terms of Bayesian random intercept and Bayesian random intercept and slope model for differentiation purposes. The posterior distributions of the cluster-specific intercepts and slopes in Bayesian multilevel propensity score models can be directly obtained and summarized. The posterior propensity scores for each subject can then be calculated based on the posterior samples of propensity score model parameters, which forms a posterior distribution of the propensity score.

In the second step, for each posterior draw of the propensity score, a treatment effect estimate and a variance estimate are obtained. Two kinds of outcome models can be applied here that provide different types of estimates conditional on the posterior propensity scores. When a Bayesian outcome model was used, the treatment effect and uncertainty estimates based on each posterior draw of the propensity score would be the conditional posterior mean or expected a posteriori (EAP) and conditional posterior standard deviation (SD) of the treatment effect. I denote this method as the *two-step full Bayesian multilevel propensity score approach*. The benefit of the full Bayesian approach is that prior information can be encoded in both the propensity score equation and the outcome equation. However, the full Bayesian approach suffers from long computing time, especially for the optimal full matching method with Bayesian multilevel outcome model. Thus, the other kind of outcome model, a conventional outcome model, can be utilized to provide treatment effect and variance estimates based on posterior propensity scores to approximate the conditional posterior mean and variance in a Bayesian outcome model with noninformative priors. I denote this method as the *two-step approximate Bayesian multilevel propensity score approach*. The advantage of the approximate Bayesian multilevel propensity score approach is that it incorporates the uncertainty in the propensity score into the final variance estimation and runs much faster than the full Bayesian approach. Also, in Bayesian propensity score analysis, the outcome model is often a simple linear model within matched sets with only treatment selection as the predictor (sometimes additional covariates are also included for further adjustment)

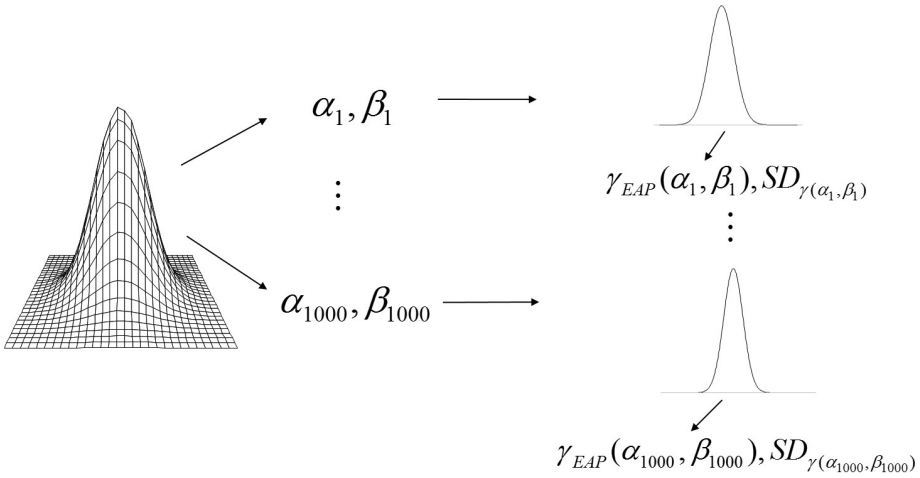

Figure 3.1: Two-step Full Bayesian Multilevel Propensity Score Approach

due to the fact that the influential covariates have been modeled in the propensity score equation. Convergence of the posterior samples usually can be achieved even with a noninformative prior. Theoretically, a Bayesian outcome model with noninformative priors provides similar results to the conventional outcome model. The graphical illustrations of the full and approximate Bayesian multilevel propensity score approaches were shown in Figures 3.1 and 3.2, respectively.

In this dissertation, I evaluated the two-step approximate Bayesian multilevel propensity score approach in the simulation studies and investigated both the approximate and full Bayesian multilevel propensity score approaches in the case study. I view these two Bayesian multilevel propensity score approaches as two variations of the proposed Bayesian multilevel propensity score approach as they only differ in the types of the outcome model and both account for the uncertainty in the propensity score. The flowchart in Figure 3.3 outlined the analytic procedure of the proposed Bayesian multilevel propensity score approach.

If there were 1000 posterior draws of the propensity score in the first step, there

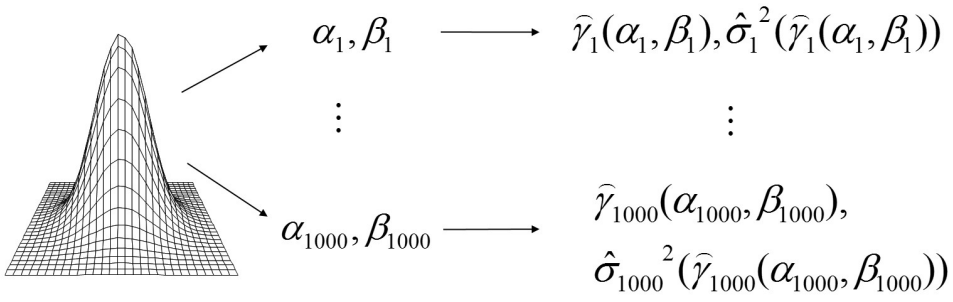

Figure 3.2: Two-step Approximate Bayesian Multilevel Propensity Score Approach

would be 1000 treatment effect estimates and variance estimates in the second step, either posterior means or EAP and posterior standard deviations of the treatment effect from a Bayesian outcome model, or treatment effect and variance estimates form a conventional outcome model. The treatment effect estimate within each cluster for within-cluster matching and the final treatment effect estimate for across-cluster matching would be the average of these 1000 posterior-based treatment effect estimates, while the variance estimate within each cluster for the within-cluster matching and the final variance estimate for the across-cluster matching would be the average variance estimates across 1000 posterior draws plus the variation of the 1000 treatment effect estimates, which accounts for uncertainty in the propensity score. The cluster-specific treatment effect and variance estimates then need to be pooled together for the within-cluster matching strategy. More details with regard to the treatment/causal effect and variance estimation are presented in the following section 3.2.

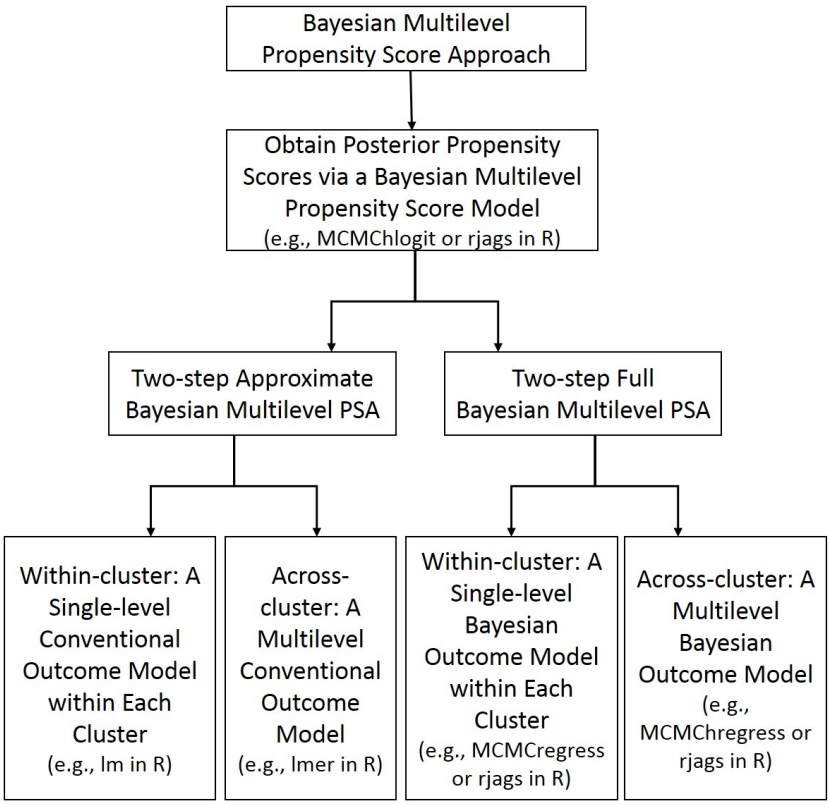

Figure 3.3: Flowchart of the Bayesian Multilevel Propensity Score Approach

3.2 Casual Effect and Variance Estimation

In multilevel observational studies, individuals can be matched within each cluster to approximate blocked randomized experiment or matched across clusters to borrow information from individuals in other clusters. The corresponding causal effect and variance estimators for within-cluster matching and across-cluster matching are slightly different.

After obtaining the posterior propensity scores from a Bayesian multilevel propensity score model, for the within-cluster matching strategy, the treatment effect estimate is obtained within each cluster and no cluster-level variation needs to be considered in this case. In this dissertation, a single-level outcome model

was fit to obtain the propensity score adjusted treatment effect estimate within each cluster. As the variances of the treatment effect estimates in different clusters are often heterogeneous, I pooled the cluster-specific treatment effect estimates together using the inverse variance of each cluster's treatment effect estimate as the weight (Hartung et al., 2008).

The cluster-specific treatment effect and variance estimates for the two-step full Bayesian approach could be obtained in the same way as the estimates in Kaplan and Chen (2012) in equation 1.5 and 1.6, respectively, but with propensity score parameters obtained from a Bayesian multilevel propensity score model. The cluster-specific treatment effect and variance estimates for the approximate Bayesian approach can also be obtained similarly, with the conditional posterior mean $\mu_{\gamma(\eta_m)}$ and posterior variance $\sigma^2_{\gamma(\eta_m)}$ based on the m^{th} posterior draw of the propensity score in equation 1.6 replaced by the treatment effect and variance estimates obtained from a conventional outcome model within each cluster. Denote the cluster-specific treatment effect and variance estimate as γ_j and σ^2_j for cluster j, respectively. Then the final pooled treatment effect estimate across clusters for the within-cluster matching strategy would be $\frac{\sum_j \gamma_j/\sigma^2_j}{\sum_j 1/\sigma^2_j}$ and the final pooled variance estimate would be $\frac{1}{\sum_j 1/\sigma^2_j}$. The inverse-variance weighted average was shown to have the least variance among all weighted averages (Hartung et al., 2008). If the variances of different clusters are all equal, then the inverse-variance weighted average becomes the simple average.

For the across-cluster matching strategy, the treatment effect is estimated in matched sets across clusters. A multilevel outcome model is fit within each matched group across clusters. The final treatment effect and variance estimates for the full Bayesian approach are the same as the estimates in Kaplan and Chen (2012) in equation 1.5 and 1.6, respectively, but with the posterior propensity scores and conditional posterior mean and variance of the treatment effect obtained from a Bayesian multilevel propensity score model and a Bayesian multilevel outcome model, respectively. The final treatment effect and variance estimates for the approximate Bayesian approach can be obtained similarly to the estimates in 1.5 and 1.6, with the conditional posterior mean $\mu_{\gamma(\eta_m)}$ and conditional posterior variance

$\sigma^2_{\gamma(\eta_m)}$ of the treatment effect based on the m^{th} posterior draw of the propensity score replaced by the treatment effect and variance estimates obtained from a conventional outcome model. No pooling across schools is needed for the across-cluster matching strategy.

4 DESIGN AND RESULTS OF SIMULATION STUDIES

4.1 Design of Simulation Studies

This dissertation conducted two comprehensive simulation studies to examine the performance of the proposed Bayesian propensity score approach for multilevel observational studies. For both simulation studies, data were generated as follows:

1. Independently generate random variables X_1 and X_2 as two level-one covariates and W_1 and W_2 as two level-two covariates such as

$$\begin{aligned} X_1 &\sim N(0, 0.5), \\ X_2 &\sim N(1, 1.5), \\ W_1 &\sim N(1, 1), \\ W2 &\sim N(0, 1). \end{aligned}$$

   The sample size of each cluster n is set as 100 and there are G =20 clusters.

2. Generate the level-two residuals for the random intercept (u_0) and two random slopes (u_1 and u_2) with mean 0 and the covariance matrix Σ_u =

$$\begin{Bmatrix} .25 & .15 & .225 \\ .15 & .36 & .27 \\ .225 & .27 & .81 \end{Bmatrix}$$

3. Generate the level-one residual $\epsilon \sim N(0, 0.1)$.

4. Obtain the true propensity score $e(X)$ based on the propensity score model

below, where $i = 1, \ldots, n$ and $j = 1, \ldots, G$:

$$\begin{aligned}
&Logit[e(X)] = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij};\\
&\beta_{0j} = 0.1 + 0.1W1 + 0.2W2 + u_{0j};\\
&\beta_{1j} = 0.2 + 0.2W1 + 0.3W2 + u_{1j};\\
&\beta_{2j} = 0.1 + 0.2W1 + 0.4W2 + u_{2j}.
\end{aligned}$$

5. Generate the treatment assignment $T \sim Bernoulli[e(X)]$.

6. Obtain the outcome Y based on the outcome model below, where the true treatment effect is set as 1.5:

$$\begin{aligned}
&Y_{ij} = \gamma_{0j} + 1.5T + \gamma_{1j}X_{1ij} + \gamma_{2j}X_{2ij} + \epsilon_{ij};\\
&\gamma_{0j} = 0.1 + 0.2W1 + 0.1W2 + u_{0j};\\
&\gamma_{1j} = -0.1 + 0.1W1 - 0.2W2 + u_{1j};\\
&\gamma_{2j} = 0.3 + 0.2W1 + 0.4W2 + u_{2j}.
\end{aligned}$$

7. Replicate step 1 to step 6 for 500 times.

In terms of the propensity score estimation, in addition to a random intercept model with all the fixed slopes and a random intercept and slope model, a single-level model with individual- and cluster- level covariates and a fixed effect model with dummy-coded school identities as covariates are also evaluated. The R program *arm* (Gelman et al., 2011) was utilized for fitting Bayesian single-level and fixed effect propensity score models in the simulation studies. For the treatment effect estimation, in addition to the random intercept model, a single-level model is also fit for comparison purposes.

This dissertation examined the optimal full matching method and regression adjustment approach in both simulation studies. For the optimal full matching method, subjects were matched into different subgroups based on their posterior propensity scores to minimize the total distance in the propensity scores between

subjects in different treatment conditions. The subgroup membership was then included as dummy variables in the outcome model along with the treatment assignment to obtain the final treatment effect. With regard to the regression adjustment approach, I included the linear, quadratic and cubic terms of the logit propensity scores as covariates in the outcome model to adjust for selection bias. For comparison purposes, the unadjusted treatment effect estimates in various simulation conditions were also evaluated via a Bayesian simple linear regression model with the treatment selection as the only predictor.

4.2 Design of Simulation Study 1

In Simulation Study 1, individuals are matched within each cluster to approximate blocked randomized trials. This study examined the performance of the Bayesian multilevel propensity score approach with different sample sizes at both individual level and cluster level. The influence of different prior information on the treatment effect and variance estimation was also investigated. For Simulation Study 1, the raw intra-class correlations range from 0.24 to 0.67 with a mean of 0.50 and median of 0.51 across 500 data replications.

The sample size within each cluster varied at $n = 100$ and $n = 250$, while the number of clusters varied at $G = 20$ and $G = 50$. Noninformative prior and informative prior in the propensity score models were compared under the case of 20 clusters with 100 individuals in each cluster. Noninformative prior were utilized for other sample size conditions. For noninformative prior, this study utilized the improper uniform prior with zero prior mean and infinite prior variance to allow the data to direct the estimates. For the informative prior, the generating coefficients of the fixed effects shown in step 4 of the section 4.1 were utilized as prior mean and 1 was chosen as prior variance for the Bayesian propensity score models to represent relatively precise prior information on the model parameters. For Bayesian multilevel propensity score models with random intercept and random intercept and slopes, there were 10,000 MCMC iterations with 5000 burnin and a thinning interval of 10.

4.3 Results of Simulation Study 1

Results of Simulation Study 1 were shown in Table 4.1. The average unadjusted treatment effect and variation estimates across 500 replications were presented in the bottom rows of each sample size condition. In addition to that, the left two columns of numbers referred to the average treatment effect estimates over 500 replications for the optimal full matching and regression adjustment methods, respectively. The numbers on the left of the slash line in the third and fourth columns indicated the standard error estimates averaged over 500 replications, while the numbers on the right of the slash line represented the standard deviations of the treatment effect estimates among 500 replications for the optimal full matching and regression adjustment methods, respectively. The standard deviation of the treatment effect estimates here was viewed as the approximately true variation and served as a benchmark to evaluate the accuracy of the uncertainty estimate.

Overall, all the Bayesian propensity score models offered less biased treatment effect estimates compared to the unadjusted effects, except that the Bayesian single-level propensity score model provided the same treatment effect estimate under the sample size of 50 clusters and 100 individuals within each cluster (1.60). This might be because there was a large amount of cluster-level variations such that the single-level propensity score model that ignored the multilevel structure in the treatment selection performed similarly as the unadjusted single-level linear model.

With regard to model specification, results revealed that overall, propensity score models that accounted for multilevel structure in the treatment selection provided less biased treatment effect estimates and more accurate uncertainty estimates compared to single-level propensity score models that ignored the multilevel data structure for all evaluated sample size conditions. Single-level propensity score models severely underestimated the variance of the treatment effect estimates. Particularly, the true variation among the treatment effect estimates from a single-level propensity score model was much larger than the true variation among the estimates from a multilevel propensity score model (e.g., 0.21 v.s. 0.08 for the

Table 4.1: Treatment Effect and Standard Error Estimates over 500 Replications in Simulation Study 1

		Avg. Trt. Esti.		Avg. S.E. / S.D. of Trt. Esti.	
Sample Size	PS Model	OPTM	Regr.	OPTM	Regr.
G=20, n=100	SL	1.64	1.63	.07/.21	.07/.21
	SL(Informative)	1.64	1.64	.07/.21	.07/.21
	FE	1.56	1.55	.07/.11	.06/.10
	RI	1.54	1.52	.04/.08	.04/.07
	RI(Informative)	1.54	1.52	.04/.09	.04/.07
	RIS	1.53	1.50	.05/.05	.04/.02
	RIS(Informative)	1.52	1.50	.05/.04	.04/.02
	Unadjusted	1.66		.08/.37	
G=50, n=100	SL	1.60	1.60	.04/.22	.04/.21
	FE	1.53	1.53	.04/.11	.04/.08
	RI	1.52	1.51	.03/.09	.02/.05
	RIS	1.53	1.52	.04/.10	.03/.07
	Unadjusted	1.60		.05/.26	
G=20, n=250	SL	1.56	1.56	.05/.17	.04/.17
	FE	1.52	1.52	.04/.08	.04/.08
	RI	1.51	1.51	.03/.07	.02/.07
	RIS	1.51	1.50	.03/.03	.03/.01
	Unadjusted	1.58		.05/.35	

*N*ote. Avg.: Average; S.E.: Standard Error; S.D.: Standard Deviation; Trt. Esti.: Treatment Effect Estimate; OPTM: Optimal full matching; Regr.: Regression; SL: Single-level Model; FE: Fixed-effects Model with School Dummies; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

RI model and 0.05 for the RIS model under the condition of G=20, n=100 for the optimal full matching method), indicating that a random intercept (and slope) model accounted for the variations in random intercept (and random slopes) well such that its(their) treatment effect estimates across 500 replications were more precise compared to the single-level model estimates. Among propensity score models that took into account cluster-level variations, random intercept and slope models offered the least biased treatment effect estimate and the most accurate uncertainty estimate.

In terms of sample size, results indicated that compared to the case of 20 clusters with 100 subjects in each, the increase in the within-cluster sample size from 100 subjects to 250 subjects improved the treatment effect estimates for all kinds of evaluated propensity score models, while the increase in the number of clusters from 20 to 50 clusters did not show as much advantage. With 50 clusters, treatment effect estimates were slightly less biased for the single-level, fixed-effects and random intercept propensity score model, but very similar for the random intercept and slope model compared to the estimates with 20 clusters. That is, for multilevel observational studies, in order to achieve good within-cluster matching to recover the treatment effect in the blocked randomized trials, the increase in the within-cluster sample size would reduce more selection bias than the increase in the number of clusters given the same total sample size.

Results also indicated that informative prior on the propensity score model did not have substantial influence and provided very similar treatment effect and uncertainty estimate compared to those with noninformative uniform priors. This result was consistent with Kaplan and Chen (2012), which showed that prior information on the propensity score model had less influence on the treatment effect estimate compared to prior information on the outcome model.

4.4 Design of Simulation Study 2

Simulation Study 1 above focused on the evaluation of sample sizes and prior information. To further study the performance of the proposed approach on different

matching strategies and intra-class correlations, I conducted Simulation Study 2 and utilized 500 randomly generated data sets, each having 20 clusters with 100 subjects in each cluster. The investigated matching strategies include within-cluster matching and across-cluster matching. For the intra-class correlation, I examined the effects of medium intra-class correlation and low intra-class correlation with mean 0.5 and 0.06, respectively, across 500 replications.

4.5 Results of Simulation Study 2

Results of Simulation Study 2 are presented in Table 4.2 and Table 4.3 for medium intra-class correlation and low intra-class correlation, respectively. The corresponding average unadjusted treatment effect and variation estimates over 500 replications were presented in the bottom rows of Table 4.2 and Table 4.3, respectively.

Results in Table 4.2 showed that similar to Simulation Study 1, propensity score models that accounted for multilevel treatment selection structure offered less biased treatment effect estimates and more accurate variation estimates compared to the single-level propensity score model for both within- and across-cluster matching. Single-level propensity score, fixed effect model with school dummies and random intercept model all underestimated the uncertainty in treatment effect estimates, while the random intercept and slope propensity score model provided the most accurate standard error estimates.

From Table 4.3, results indicated that when intra-class correlation on the outcome was low, overall, multilevel propensity score models using the across-cluster matching strategy offered less biased treatment effect estimates. Propensity score models that allowed both intercept and slopes to vary across clusters produced more accurate uncertainty estimates for both within- and across-cluster matching. No matter for the data with medium or with low intra-class correlations, single-level and random intercept outcome model yielded very similar results.

Overall, from both Table 4.2 and Table 4.3, propensity score methods provided less biased treatment effect estimates compared to the unadjusted estimates, except that single-level propensity score models performed very similarly with the un-

Table 4.2: Performance metrics averaged over 500 replications. Sample size $G = 20$ and $n = 100$. ICC $\sim$ (.24, .67) with a mean of .50 and median of .51

	PS Model	Outcome Model	Avg. Trt. Esti.		Avg. S.E. / S.D. of Trt. Esti.	
			OPTM	Regr.	OPTM	Regr.
Within-Cluster	SL	SL	1.64	1.63	.07/.21	.07/.21
	FE	SL	1.56	1.55	.07/.11	.06/.10
	RI	SL	1.54	1.52	.04/.08	.04/.07
	RIS	SL	1.53	1.50	.05/.05	.04/.02
Across-Cluster	SL	SL	1.66	1.65	.10/.34	.09/.35
		RI	1.63	1.64	.08/.22	.08/.24
	FE	SL	1.60	1.61	.17/.28	.16/.28
		RI	1.58	1.59	.11/.18	.11/.20
	RI	SL	1.50	1.50	.10/.15	.09/.15
		RI	1.51	1.51	.06/.11	.06/.12
	RIS	SL	1.51	1.51	.13/.09	.11/.05
		RI	1.51	1.51	.08/.06	.07/.08
Unadjusted			1.66		.08/.37	

*N*ote. Avg.: Average; S.E.: Standard Error; S.D.: Standard Deviation; Trt. Esti.: Treatment Effect Estimate; OPTM: Optimal full matching; Regr.: Regression; SL: Single-level Model; FE: Fixed-effects Model with School Dummies; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

adjusted single-level linear model. As both the propensity score and the outcome were generated using the random intercept and slope model in the simulation studies, this result indicated that in the multilevel settings, single-level propensity score model might even not outperform the unadjusted simple linear regression. Similar to Simulation Study 1, the standard deviation of the treatment effect estimates across 500 replications, i.e., the approximate true variation dropped as the variations in intercept and slopes were taken into account (e.g., 1.95 v.s. 0.51 for the optimal matching method for within-cluster matching in Table 4.3). These results implied the importance of accounting for multilevel structure of the treatment

Table 4.3: Performance metrics averaged over 500 replications. Sample size G = 20 and n = 100. ICC ~ (.01, .13) with a mean of .06 and median of .05

	PS Model	Outcome Model	Avg. Trt. Esti.		Avg. S.E. / S.D. of Trt. Esti.	
			OPTM	Regr.	OPTM	Regr.
Within-Cluster	SL	SL	2.40	2.36	.41/1.95	.40/1.94
	FE	SL	1.77	1.67	.50/1.29	.49/1.26
	RI	SL	1.78	1.62	.29/1.05	.22/.93
	RIS	SL	1.72	1.53	.46/.51	.34/.21
Across-Cluster	SL	SL	2.36	2.35	.40/2.08	.37/2.07
		RI	2.38	2.36	.41/2.09	.37/2.06
	FE	SL	1.91	1.91	.71/1.84	.62/1.79
		RI	1.91	1.95	.71/1.82	.60/1.72
	RI	SL	1.53	1.54	.35/1.38	.31/1.32
		RI	1.53	1.56	.34/1.37	.29/1.28
	RIS	SL	1.56	1.53	.57/.28	.46/.42
		RI	1.56	1.51	.56/.28	.44/.65
Unadjusted			2.37		.26/1.84	

*N*ote. Avg.: Average; S.E.: Standard Error; S.D.: Standard Deviation; Trt. Esti.: Treatment Effect Estimate; OPTM: Optimal full matching; Regr.: Regression; SL: Single-level Model; FE: Fixed-effects Model with School Dummies; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

selection in multilevel observational studies.

4.6 Summary of Results in Simulation Studies

Overall, under the conditions of the simulation studies, I found that:

1. The proposed Bayesian multilevel propensity score approach offered less biased treatment effect estimate and more accurate uncertainty estimate compared to Bayesian propensity score models that ignored the multilevel structure.

2. For the within-cluster matching strategy, the increase in the individual-level sample size offered better treatment effect and variation estimates compared to the increase in the number of clusters given the same total sample size.

3. When the intra-class correlation was medium and the cluster-specific variation was properly accounted for, within-cluster matching and across-cluster matching provided similar results, but the within-cluster matching strategy offered the most accurate uncertainty estimate under the optimal matching method.

4. When the intra-class correlation was low, the across-cluster matching strategy provided less biased treatment effect estimates.

5. Prior in the propensity score model had little impact on the treatment effect estimation.

6. When treatment selection has a multilevel structure, single-level propensity score model might not be able to effectively reduce the initial selection bias.

In general, Bayesian random intercept and slope propensity score model for within-cluster matching strategy is recommended for multilevel observational studies based on its overall performance of treatment effect and variance estimates. When there is little evidence of omitted cluster-level covariates, Bayesian multilevel model with across-cluster matching may provide as good treatment effect and variation estimates.

5 DESIGN AND RESULTS OF THE CASE STUDY

5.1 Design of the Case Study

5.1.1 Data

In order to investigate the practical performance of the proposed Bayesian multilevel propensity score procedure, I conducted a case study using the real data from the Early Childhood Longitudinal Study Kindergarten cohort of 1998 (ECLS-K) (NCES, 2001). The ECLS-K is a nationally representative longitudinal sample providing comprehensive information from children, parents, teachers and schools. It follows the same students from kindergarten through the eighth grade and offers an excellent resources for conducting research in children's educational development.

Whether grade retention provides value for at-risk children in their later development has been attracting educational researchers' attention for a long time. Though research showed mixed results with regard to the impact of early grade retention on children's academic achievement and psychosocial skills (e.g., Hong and Yu, 2007, 2008; Wu et al., 2010), fourteen states and the District of Columbia have enacted policies requiring students who do not show basic reading proficiency by the end of third grade to be retained and offered with remediating services (Rose, 2012). Using the ECLS-K data set, this dissertation investigates the causal effect of kindergarten retention on children's later reading achievement at the end of first grade (Spring of 2000) to see whether the retainees benefit from the retention treatment compared to their promoted counterparts through the Bayesian multilevel propensity score approach.

Among schools that allowed for kindergarten retention, there are 13469 students in 1064 schools that the school identities and students' retention status were clearly specified at the end of the kindergarten year. These include 463 retained students and 13006 promoted students. I calculated the sample retention rates by taking the proportion of the sampled kindergartners retained at the end of kindergarten

year to the total sampled children in the pre-treatment kindergarten year in each school, which can be viewed as an unbiased estimate of each school's kindergarten retention rate (Hong and Raudenbush, 2006). The retention rates vary from 0 to unfeasible 1 with a weighted mean at 0.03. Among 1064 schools, there are 824 schools that have zero student being retained or being promoted in the sample, which makes the within-school matching impossible. In the rest 240 schools that have at least one student in both treatment conditions, 144 schools have only one promoted student or one retained student in the sample and this one student will for sure be matched with all students in the other treatment condition for the within-school matching method. The within-school retention effect for these schools are then simply the mean difference between two treatment conditions. This implied the practical limitation of the within-cluster matching that it may be hard to find close matches or even any match within each cluster. Across-cluster matching could still be employed in this case as students are allowed to match across clusters. To facilitate the within-cluster matching, schools with fewer than two retained or promoted students were removed. This yielded 1531 students in 96 schools, including 1254 promoted students and 277 retained students, which constituted the analytic sample of this case study.

The schools in the analytic sample have retention rates ranging from 0.09 to 0.50, which are all higher than the national average of 0.05 in 1995 (Zill et al., 1997). Thus, the schools in the analytic sample have relatively high retention rates. I utilized the same covariates as those were used in the propensity score model for high-retention schools in Hong and Raudenbush (2006) for obtaining posterior propensity scores. These include eleven student-level covariates and two school-level covariates. Specifically, student-level covariates used in the case study are reading item response theory (IRT) scale score collected at the spring kindergarten, general knowledge academic rating scale (ARS) score measured at the fall and spring kindergarten, parent's rating of child approaches to learning, teacher's report of child in the lowest reading group at the spring kindergarten, child's age at kindergarten entry, child's gender, parent's report of paying tuition for child education at the spring kindergarten, parent's report of child with disability at

the fall kindergarten, proportion of boys in the fall kindergarten class and teacher report of emphasizing importance of home-assisted kindergarten learning at the fall kindergarten. School-level covariates include principal report of teacher union and administration working together and school aggregated reading IRT scale score collected at the spring kindergarten. Therefore, there are thirteen covariates in total.

Missing data were handled via the multivariate imputation by chained equations using the R function *mice* in the *mice* package (Van Buuren and Groothuis-Oudshoorn, 2011) with one iteration. Student-level and school-level variables were imputed separately and then combined together. The diagnostics were performed as default and the chains converged. The observed intra-class correlation of the reading scores at the end of the first grade in the analytic sample is 0.21, which indicates that there is a fair amount of within-school dependences in students' reading scores.

5.1.2 Design

This case study investigated the use of Bayesian multilevel propensity score approach with random intercept propensity score model and random intercept and slope propensity score model to evaluate the kindergarten retention effect. The slopes of child's general knowledge at the spring kindergarten and teacher's report of emphasizing importance of home-assisted kindergarten learning at the fall kindergarten were allowed to vary across schools and to be correlated with the random intercept for the Bayesian random intercept and slope model. For comparison purposes, a single-level propensity score model with all the student-level and school-level covariates were also evaluated. Optimal full matching method was utilized for matching students with similar covariate background. For within-cluster matching, a single-level linear model was adopted as the outcome model, while for across-cluster matching, the outcome model employed a single-level or a multilevel linear model for comparison purposes. Both the two-step approximate and full Bayesian multilevel propensity score methods were examined, which adopt

a conventional outcome model and a Bayesian outcome model, respectively.

All the analyses were conducted in the R (R Development Core Team, 2012) environment. In particular, the R package *rjags* (Plummer, 2014) was utilized for obtaining posterior samples of the propensity score. I used 100,000 iterations with a thinning interval 100, burnin 10,000 and adaptive steps 10,000 for Bayesian propensity score models. The R package *optmatch* (Hansen and Klopfer, 2006) was used for implementing the optimal full matching method. The R program *lmer* in the *lme4* (Bates et al., 2013) package and the R package *rjags* (Plummer, 2014) were chosen to fit the conventional multilevel outcome model and the Bayesian multilevel outcome model, respectively. There were 10,000 iterations with a thinning interval 10 and burnin 5000 for the Bayesian multilevel outcome model. The R program *MCMCregress* in the *MCMCpack* (Martin et al., 2011) was utilized for fitting a Bayesian single-level outcome model, with 1000 iterations, thinning of 1 and 1000 burnin. All Bayesian model results were based on one MCMC chain. Noninformative priors were chosen for all Bayesian propensity score and outcome models in the case study. The R code of the two-step Bayesian multilevel propensity score approach for the case study was presented in the Appendix.

This study also evaluated covariate balance performance of the Bayesian multilevel propensity score approach. The continuous covariates and categorical covariates were evaluated separately. The balance indices used here were the standardized mean/proportion difference (Cohen's d) (Cohen, 1988) and variance ratio for each continuous covariate/each level of categorical covariates between retained students and promoted students. Specifically, the standardized mean difference for a continuous covariate was obtained by

$$B_1 = (\bar{x_t} - \bar{x_c})/\sqrt{(s_t^2 + s_c^2)/2}, \tag{5.1}$$

where $\bar{x_t}$ and $\bar{x_c}$ are the sample mean of each covariate in retained and promoted groups, respectively, and s_t^2 and s_c^2 are corresponding sample variances. The variance ratio for a continuous covariate, R_1, is defined as s_t^2/s_c^2. All the categorical covariates were dummy coded.

For each categorical level, this study evaluated the standardized difference in proportions between different treatment conditions, consistent with Harder et al. (2010). The standardized proportion difference was calculated by

$$B_2 = (\hat{p}_t - \hat{p}_c)/\sqrt{[\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_c)]/2}, \tag{5.2}$$

where $\hat{p}_t$ and $\hat{p}_c$ are proportions of participants in the retained group and promoted group, respectively, for a specific level of a categorical covariate. The variance ratio for a certain categorical level, R_2, was calculated by $\hat{p}_t(1-\hat{p}_t)/\hat{p}_c(1-\hat{p}_c)$.

In addition to the above point estimates, a Bayesian propensity score approach provides 95% posterior probability intervals (PPI) of the standardized mean/proportion difference and the variance ratio based on the posterior propensity scores. For each set of posterior propensity scores, a point estimate of each balance index can be obtained. As there were 1000 posterior draws of the propensity score in this study, a distribution of the balance indices was formed. I extracted the mean of this posterior distribution as the point estimate of the covariate balance and the 2.5^{th} and 97.5^{th} percentiles to obtain the corresponding 95% posterior probability interval for each covariate/categorical level.

5.2 Results of the Case Study

5.2.1 Convergence and Stationarity Check

The convergence and autocorrelation of the MCMC chains in the Bayesian propensity score models were first evaluated. The trace plots indicated that all the posterior propensity scores in the Bayesian single-level propensity score model converged well, while most posterior propensity scores for the Bayesian multilevel propensity score models achieved convergence. The trace plots and posterior density plots of selected posterior propensity scores obtained by Bayesian single-level, Bayesian random intercept and Bayesian random intercept and slope propensity score models were shown in Figures 5.1, 5.2 and 5.3, respectively.

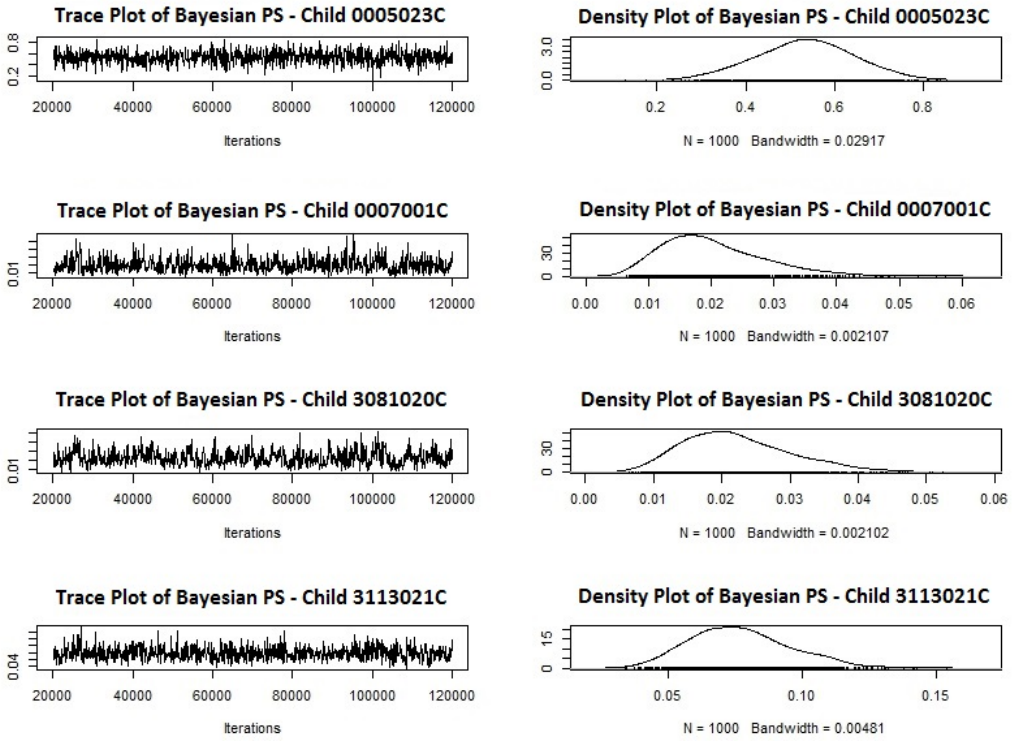

Figure 5.1: Trace and Posterior Density Plots of Selected Posterior Propensity Scores for Bayesian Single-Level Propensity Score Model in the Case Study

Overall, there was no undesirable amount of autocorrelation existing in the MCMC chains. The autocorrelation plots of selected posterior propensity scores obtained by Bayesian single-level, Bayesian random intercept and Bayesian random intercept and slope propensity score models were presented in Figures 5.4, 5.5 and 5.6, respectively. The same four children from four different schools were chosen to illustrate the convergence and autocorrelation of the MCMC chains in Bayesian propensity score models with the child identity numbers shown at the top of the plots.

The Geweke diagnostic (Geweke, 1992) and Heidelberg-Welch (Heidelberger and Welch, 1983) convergence test were also conducted to further evaluate the

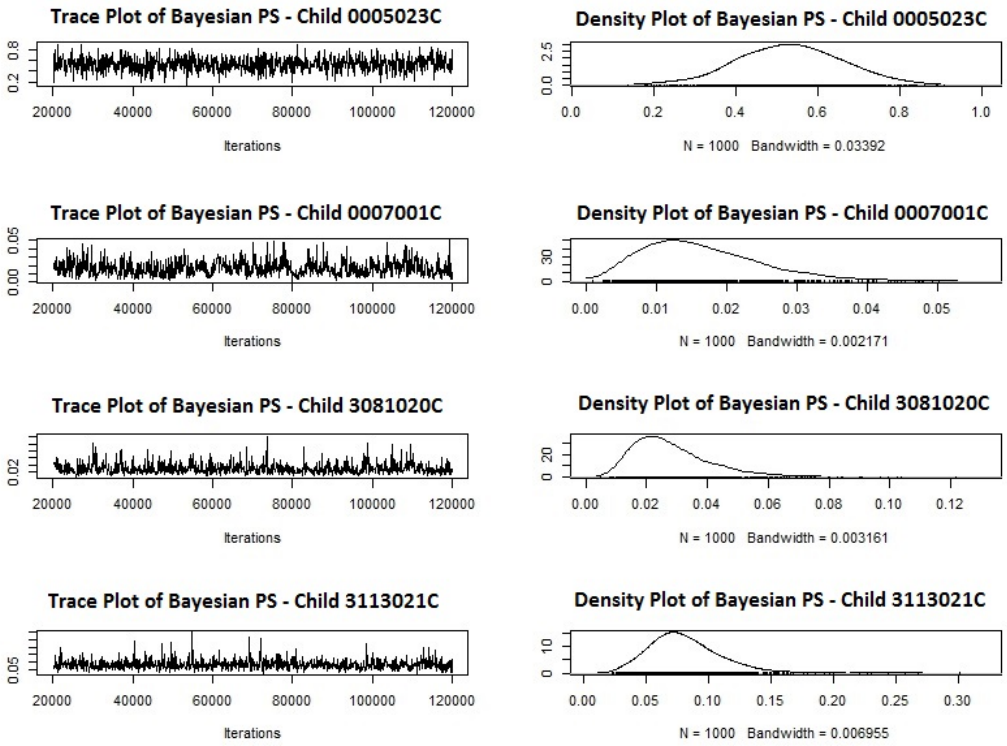

Figure 5.2: Trace and Posterior Density Plots of Selected Posterior Propensity Scores for Bayesian Random Intercept Propensity Score Model in the Case Study

convergence and stationarity of the MCMC chains in Bayesian single-level and multilevel propensity score models. The Geweke diagnostic (Geweke, 1992) examines the convergence of a Markov chain based on an equality test of the means of the first and last part of the Markov chain (in this dissertation, the first 10% and the last 50% were used). If the posterior samples were from a stationary chain, then the two means are equal and the Geweke's statistic has an asymptotically standard normal distribution. Taking a type I error rate of 0.05, any Geweke's statistic no more than 1.96 would indicate good convergence.

The Heidelberg-Welch diagnostic (Heidelberger and Welch, 1983) tests the null hypothesis that the posterior sample comes from a stationary distribution. The test

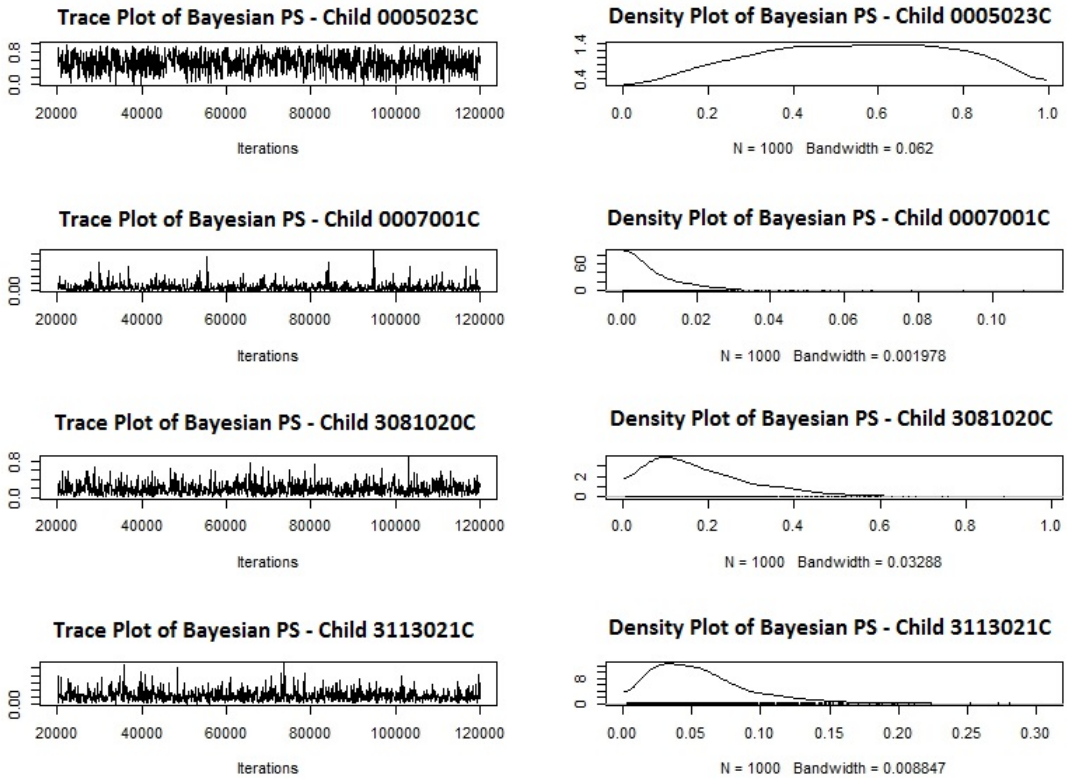

Figure 5.3: Trace and Posterior Density Plots of Selected Posterior Propensity Scores for Bayesian Random Intercept and Slope Propensity Score Model in the Case Study

is applied in succession, firstly to the whole chain, then after discarding the first 10%, 20%, . . ., of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded. The former produces a "pass" outcome, while the later yields a "failure" outcome of the stationarity test.

The convergence diagnostic results were tabulated in 5.1. Overall, all the Bayesian propensity score models achieved acceptable convergence. The mean and median Geweke's Z scores of 1531 children's posterior propensity scores were much smaller than 1.96 for all the Bayesian propensity score models. The percentages of convergent and stationary posterior propensity scores based on the Geweke's

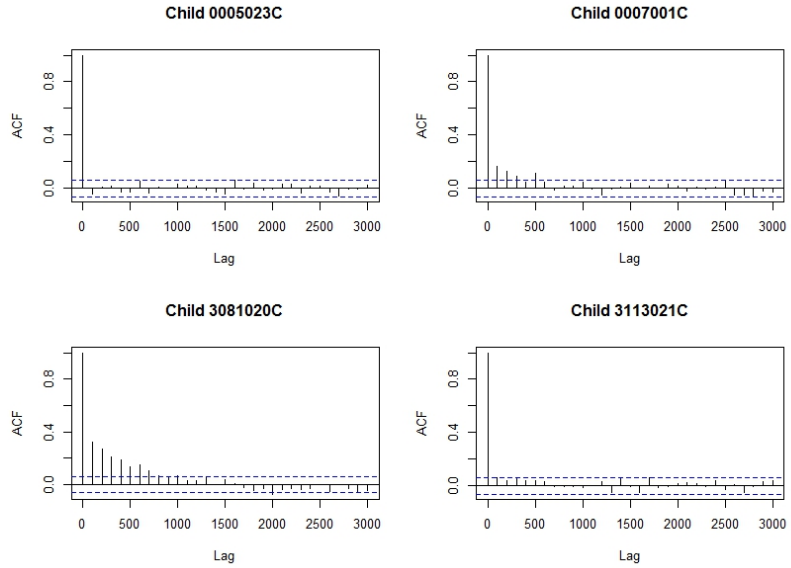

Figure 5.4: Autocorrelation Plots of Selected Posterior Propensity Scores for Bayesian Single-level Propensity Score Model in the Case Study

Table 5.1: Convergence Diagnostics of Bayesian Propensity Score Models in the Case Study

PS	Geweke's Z Score		Heidelberg-Welch diagnostic
Model	Mean/Median	% of Z <= 1.96	% of Passed
SL	.01/-.05	100 %	96.28%
RI	.42/.32	93.34 %	97.00%
RIS	.01/.04	95.82 %	87.00%

*N*ote. SL: Single-level Model; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

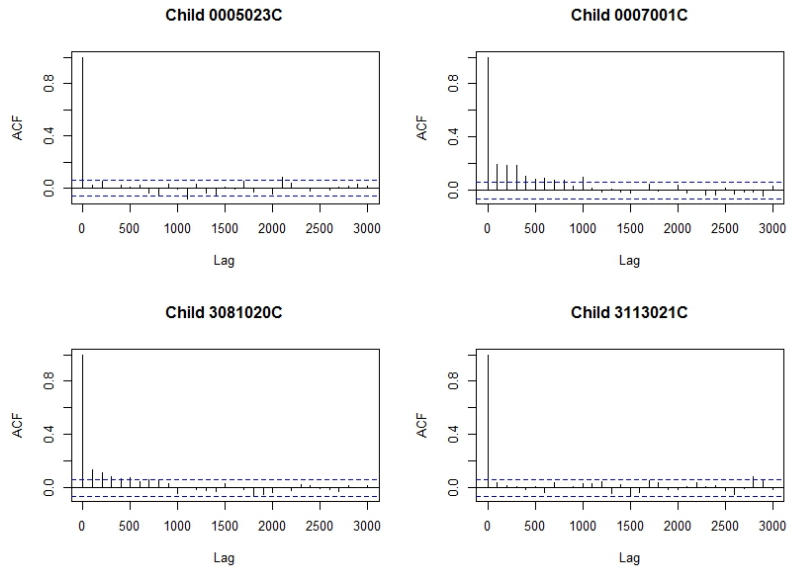

Figure 5.5: Autocorrelation Plots of Selected Posterior Propensity Scores for Bayesian Random Intercept Propensity Score Model in the Case Study

statistic and Heidelberg-Welch diagnostic indicated that Bayesian single-level and random intercept propensity score achieved better convergence and stationarity compared to Bayesian random intercept and slope model.

The Geweke plots for the same four children's posterior propensity scores were displayed in Figures 5.7, 5.8 and 5.9 for Bayesian single-level, Bayesian random intercept and Bayesian random intercept and slope propensity score models, respectively. The Geweke's Z scores for these four children were all below 1.96 for all three kinds of Bayesian propensity score models.

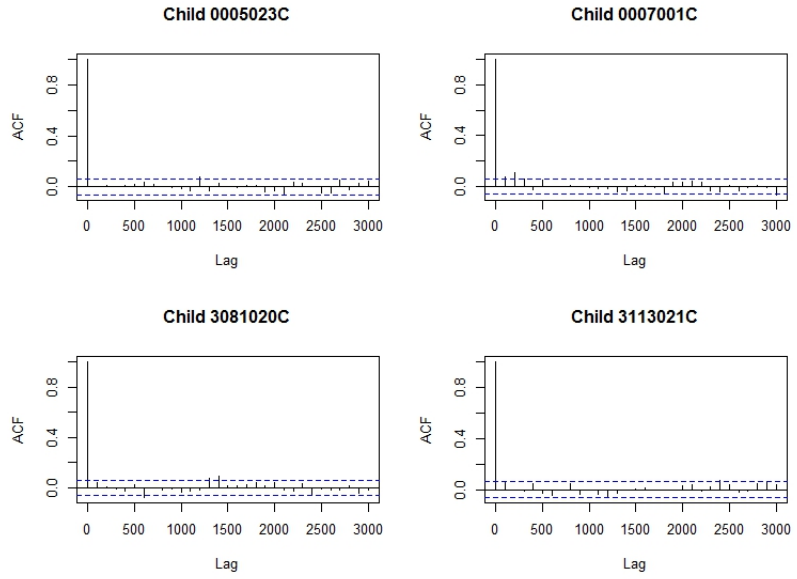

Figure 5.6: Autocorrelation Plots of Selected Posterior Propensity Scores for Bayesian Random Intercept and Slope Propensity Score Model in the Case Study

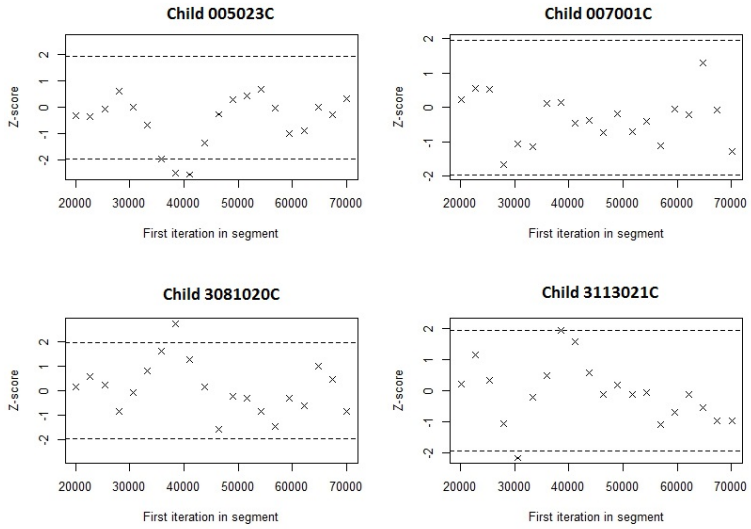

Figure 5.7: Geweke Plots of Selected Posterior Propensity Scores for Bayesian Single-level Propensity Score Model in the Case Study

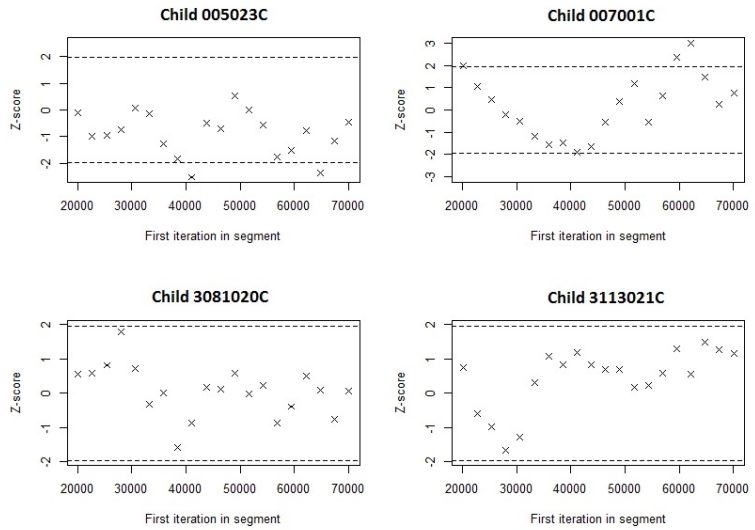

Figure 5.8: Geweke Plots of Selected Posterior Propensity Scores for Bayesian Random Intercept Propensity Score Model in the Case Study

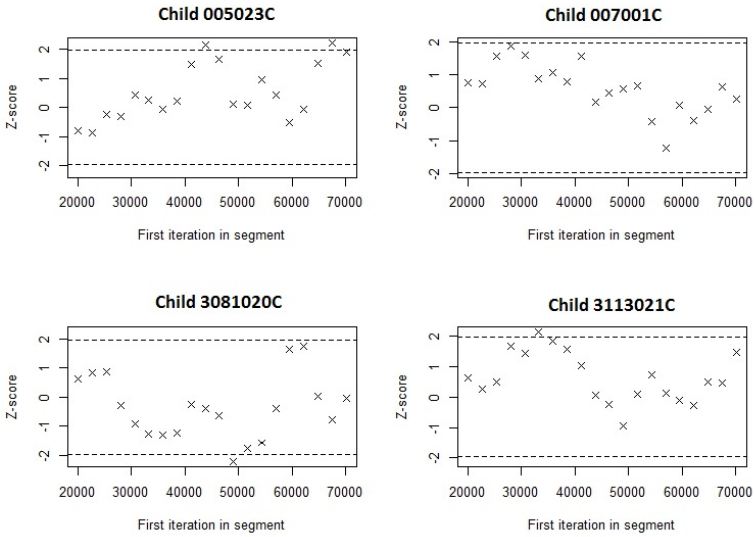

Figure 5.9: Geweke Plots of Selected Posterior Propensity Scores for Bayesian Random Intercept and Slope Propensity Score Model in the Case Study

5.2.2 Retention Effect Estimate and Covariate Balance Check

The retention effect and variance estimates of the two-step approximate and full Bayesian methods were presented in Table 5.2 and Table 5.3, respectively. For comparison purposes, a Bayesian simple linear model was fit with the treatment selection as the only predictor for children's reading achievement in the first grade to obtain the unadjusted retention effect. There were 10,000 MCMC iterations with a thinning interval of 1 and burnin of 1000 for the Bayesian simple regression. This unadjusted raw difference in reading scores served as a "negative benchmark" to check whether the Bayesian propensity score approach investigated in this study adjusted for the selection bias. I also fit a Bayesian covariate-adjusted regression with the retention status and all the thirteen covariates as the predictors. There were 10,000 MCMC iterations with a thinning interval of 1 and burnin of 5000. Noninformative priors were adopted. The results of the Bayesian simple regression and covariate-adjusted regression were displayed in the bottom rows of Table 5.2 and Table 5.3.

Overall, all the retention effect estimates with Bayesian propensity score adjustment were smaller than the unadjusted raw reading score difference between retained students and promoted students (-19.52), indicating that Bayesian propensity score approaches effectively reduced the selection bias in the retention effect estimate, which is consistent with the results in Hong and Raudenbush (2006), where retention effect estimates using propensity score stratification adjustment via across-cluster matching were much smaller than the raw retention effect. All retention effect estimates were negative, suggesting that the retained kindergartners on average scored lower in reading compared to their promoted counterparts.

The results of the two-step approximate Bayesian and the two-step full Bayesian multilevel propensity score methods paralleled each other. That is, with the same Bayesian propensity score model, the conventional outcome model and the Bayesian outcome model with noninformative priors provided very similar treatment effect and uncertainty estimates.

In terms of the matching strategy, overall, the across-cluster matching strategy

Table 5.2: Retention Effect and Standard Error (S.E.) Estimates for the Two-step Approximate Bayesian Multilevel Propensity Score Approach with a Bayesian multilevel Propensity Score Model and a Conventional Outcome Model

Matching	PS Model	Outcome Model	Retention Effect	S.E.	95% PPI
Within-Cluster	SL	SL	-16.17	.75	(-17.64, -14.70)
	RI	SL	-16.20	.75	(-17.67, -14.73)
	RIS	SL	-16.81	.76	(-18.30, -15.31)
Across-Cluster	SL	SL	-9.34	1.19	(-11.67, -7.00)
		RI	-9.04	.96	(-10.93, -7.15)
	RI	SL	-9.25	1.20	(-11.61, -6.89)
		RI	-9.00	.96	(-10.88, -7.12)
	RIS	SL	-9.58	1.34	(-12.22, -6.95)
		RI	-8.99	1.07	(-11.09, -6.89)
Unadjusted		SL	-19.52	.87	(-21.26,-17.83)
Covariate-Adjusted		SL	-9.85	.66	(-11.12, -8.57)

*N*ote. PPI: Posterior Probability Interval; SL: Single-level Model; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

offered retention effect estimates that were much smaller than the unadjusted effect and very close to the covariate-adjusted (ANCOVA) estimates (-9.85), while the within-cluster matching strategy provided retention effect estimates that were smaller, but not much, than the unadjusted effect, indicating that across-cluster matching may reduce more selection bias than the within-cluster matching in this case study. This showed a practical limitation of the within-cluster matching strategy, consistent with the findings in Thoemmes and West (2011) and Kim and Seltzer (2007). In the initial sample of the ECLS-K data set, a lot of schools have only one or no student in either treatment conditions and have been removed to facilitate the within-cluster matching. However, even with the current analytic sample, the within-school sample sizes range from 7 to 23 with the number of retained students varying from 2 to 8 and the number of promoted students varying from 4 to 21.

Table 5.3: Retention Effect and Standard Error (S.E.) Estimates for the Two-step Full Bayesian Multilevel Propensity Score Approach with a Bayesian multilevel Propensity Score Model and a Bayesian Outcome Model

Matching	PS Model	Outcome Model	Retention Effect	S.E.	95% PPI
Within-Cluster	SL	SL	-16.39	.83	(-18.02, -14.76)
	RI	SL	-16.43	.83	(-18.06, -14.80)
	RIS	SL	-16.96	.84	(-18.61, -15.31)
Across-Cluster	SL	SL	-9.33	1.20	(-11.68, -6.98)
		RI	-9.18	.96	(-11.06, -7.30)
	RI	SL	-9.24	1.21	(-11.61, -6.87)
		RI	-9.13	.96	(-11.06, -7.26)
	RIS	SL	-9.58	1.34	(-12.21, -6.95)
		RI	-9.10	1.06	(-11.18, -7.01)
Unadjusted		SL	-19.52	.87	(-21.26,-17.83)
Covariate-Adjusted		SL	-9.85	.66	(-11.12, -8.57)

*N*ote. PPI: Posterior Probability Interval; SL: Single-level Model; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

The small within-cluster sample size limited the advantage of the within-cluster matching strategy. Nonetheless, within-cluster matching approximates blocked randomized design and its performance is expected to improve as the strength of treatment selection and within-cluster sample sizes increase.

With regard to the model specification, under the same matching strategy, Bayesian single-level propensity score model and multilevel propensity score models performed similarly in terms of the retention effect and uncertainty estimates. Nevertheless, Bayesian random intercept and slope propensity score model with a single-level outcome model provided the closest retention effect estimate with the Bayesian covariate-adjusted estimate.

The average absolute standardized mean/proportion differences (Cohen's d) and variance ratio across all the covariates and categorical levels were shown in

Table 5.4. The average initial bias and raw variance ratio between the retained students and promoted students were also calculated and were presented in the bottom row of Table 5.4 (0.33 and 1.16, respectively), indicating that there is fair, but not very strong, treatment selection in the analytic sample. Results of the covariate balance after propensity score adjustments suggested that overall, all the Bayesian propensity score models greatly reduced the initial bias. The bias reduction ranged from 30.3 % to 51.5 %. Surprisingly, Bayesian single-level model and random intercept model outperformed the Bayesian random intercept and slope model on covariate balance, which might be due to the slopes that were allowed to be random did not vary greatly across schools in the analytic sample.

Table 5.4: Average Covariate Balance Performance over All Covariates/Categorical Levels in the Case Study

PS Model	Avg. Absolute Cohen's d/95% PPI	Avg. Variance Ratio /95% PPI
SL	.16/(.11, .21)	1.08/(.95, 1.22)
RI	.17/(.12, .22)	1.08/(.95, 1.21)
RIS	.23/(.16, .32)	1.11/(.97, 1.24)
	Initial Bias: .33	Raw Variance Ratio: 1.16

*N*ote. Avg.: Average; PPI: Posterior probability interval; SL: Single-level Model; RI: Random Intercept Model; RIS: Random Intercept and Slope Model.

For the within-cluster matching strategy, I also fit a Bayesian single-level propensity score model using student-level covariates combined with a conventional single-level outcome model within each school to avoid the impact of school-level omitted covariates. There were 100,000 MCMC iterations with a thinning interval of 10, burnin of 10,000 and 10,000 adaptive steps. However, the posterior propensity scores, in general, were very extreme, close to 0 or 1, for schools with very few sampled students. The trace plots of selected posterior propensity scores obtained by Bayesian single-level propensity score model were shown in Figure 5.10, which

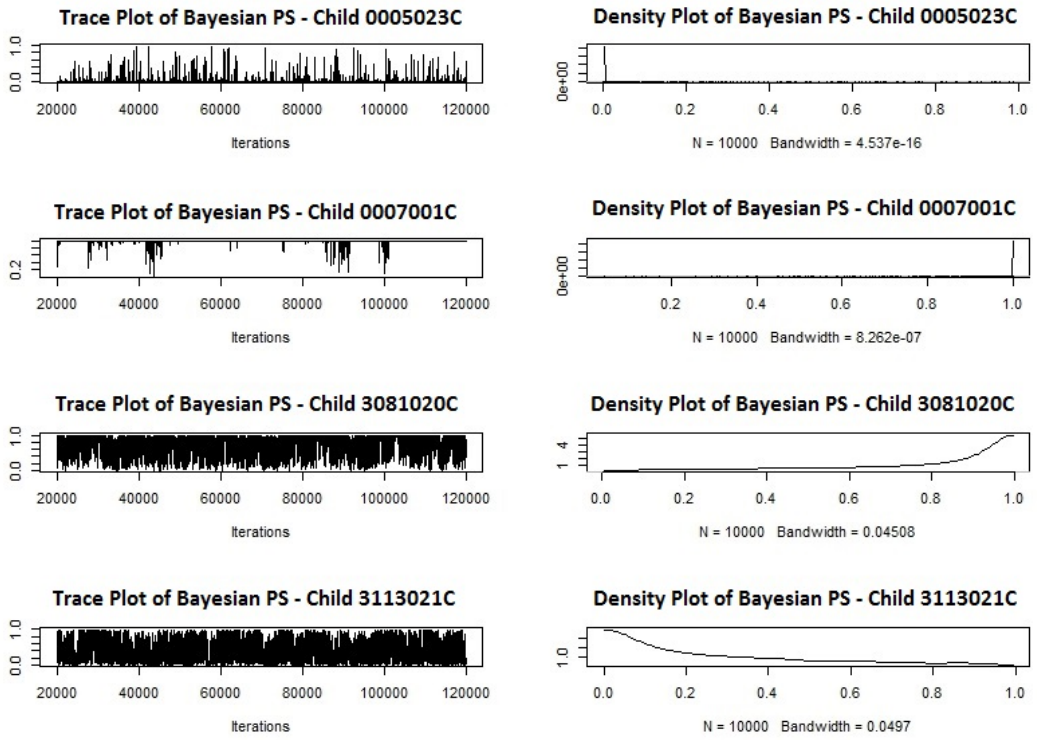

Figure 5.10: Trace and Posterior Density Plots of Selected Within-school Posterior Propensity Scores for Bayesian Single-Level Propensity Score Model in the Case Study

indicated poor convergence of some posterior propensity scores due to extremely small within-cluster sample sizes. The retention effect estimate was -20.58 with a standard error of 0.79. The retention effect estimate was even larger than the raw effect and in the opposite direction compared to the retention estimates via other propensity score models, indicating that estimating the propensity score within each school might not reduce the selection bias when within-cluster sample size was too small to facilitate close matches.

To sum up, the case study showed that the retained kindergartners in high-retention schools, on average, did not perform as well as their promoted counterparts. Bayesian single-level and multilevel propensity score models provided

smaller retention effect estimates compared to the unadjusted retention effect, consistent with the results in Hong and Raudenbush (2006), indicating that Bayesian propensity score methods may effectively reduce the selection bias. The retention effect estimates offered by the across-cluster matching strategy were closer to the covariate-adjusted results compared to the within-cluster matching strategy, suggesting that across-cluster matching might be an alternative when within-cluster matching encounters practical difficulty in facilitating close matches. The selection bias may not be effectively reduced by within-cluster matching when the overall school sizes are small and the strength of treatment selection is not high.

6 DISCUSSION AND CONCLUSION

A vast amount of literature has confirmed the effectiveness of propensity score methods in reducing selection bias in treatment effect estimation caused by non-random treatment selection/assignment in observational studies. Comparing with other treatment effect estimation methods, for example, the covariate-adjusted regression method, propensity score methods enjoy the benefit of robustness against model misspecification, ability of checking covariate balance prior to data collection, and capacity of guiding the design of observational studies. Thus, propensity score methods have been most commonly applied to make causal inference in observational studies. In the health and medical sciences alone, the numbers of published applications of propensity score analysis increased from about 87 articles per year to more than 1000 per year in the 10 years between 2004 to 2014 [1].

Various techniques of obtaining propensity score adjusted treatment effect estimates have been proposed, including propensity score stratification, weighting, matching and regression-adjustment methods. In general, the optimal full matching method is recommended because of its better alignment of comparable subjects in different treatment conditions than other matching methods (Hansen, 2004; Gu and Rosenbaum, 1993), relatively stable performance compared to the weighting method and more robust to the functional form misspecification than the regression-adjustment method. However, there is no guarantee that it performs better for a given data set (Steiner and Cook, 2013).

In addition to the treatment effect estimation, another key issue in observational study is the uncertainty estimation of the treatment effect, which is central to the statistical inference, such as confidence interval estimation and hypothesis testing. For the uncertainty estimation in propensity score analysis, various variance estimators have been proposed in the frequentist setting (Lunceford and Davidian, 2004; Abadie and Imbens, 2009), but all of them reply on large sample size in order to have close approximation to the true variance. In contrast, Bayesian propensity score

[1] Data was retrieved from http://www.ncbi.nlm.nih.gov/pubmed

methods can naturally capture the uncertainty in the propensity score and provide exact variance estimate in the Bayesian sense, i.e., posterior variance, although the posterior probability intervals might not achieve the nominal coverage rate in the frequentist repeated sampling framework (Kaplan and Chen, 2012).

A key assumption for the propensity score methods is the independence of the treatment selection and independence of potential outcomes among individuals, which is violated when data has hierarchical structure and subjects are dependent with others in the same cluster. To take into account the dependencies of treatment selection and potential outcomes, multilevel models have been proposed to model the propensity scores and/or outcomes depending on the data structure and matching strategy. Results of this dissertation showed that single-level and multilevel outcome model performed similarly once the multilevel structure of the treatment effect was accounted for.

Similar to the single-level case, uncertainty estimation of the treatment effect is central to the statistical inference in the multilevel setting. To my best knowledge, there is no available variance estimator under the conventional framework that takes into account both uncertainty in the propensity score and uncertainty in the outcome in the multilevel settings. One possible reason is the complicated variance derivation in the frequentist framework due to the dependencies caused by the multilevel data and the two-step nature of the propensity score procedure. One feasible and reliable way to obtain variance estimate is to apply Bayesian methodology to multilevel observational studies, which was investigated in this dissertation. Draper (1995) pointed out that Bayesian treatment of hierarchical models via MCMC or other sampling techniques has developed rapidly in the statistics society but has drawn little attention in many areas of social sciences, especially in education. Compared to the complexity of estimating variance in the likelihood-based multilevel logistic regression model, Bayesian hierarchical logistic regression can directly provide the uncertainty estimate once the posterior sample is obtained. In addition, the fully Bayesian hierarchical methods have been incorporated into several software packages such as WinBUGS (Spiegelhalter et al., 1999), MCMCpack (Martin et al., 2011) and rjags (Plummer, 2014) in R (R

Development Core Team, 2012), which makes the Bayesian multilevel propensity score method technically feasible.

This dissertation takes a Bayesian perspective and provides a two-step Bayesian propensity score approach for multilevel observational studies. Two variations of the approach are provided, which are two-step full Bayesian and two-step approximate Bayesian multilevel propensity score methods. Both methods account for uncertainty in the propensity score.

To sum up, the dissertation began with the notation of the potential outcomes framework for causal inference followed by the propensity score methods and multilevel modeling from both Bayesian and frequentist perspectives. The dissertation then reviewed the multilevel propensity score methods for making causal inference in multilevel observational studies and pointed out the need of developing a Bayesian multilevel propensity score approach, which is the main focus of this dissertation. The two-step Bayesian propensity score approach in the multilevel context was developed and the properties of the proposed approach were examined for different sample sizes, prior information, intra-class correlations, matching strategies and propensity score methods through two comprehensive simulation studies and one real-data case study.

Results of the simulation studies showed that the proposed Bayesian multilevel approach offered less biased treatment effect estimate and more accurate uncertainty estimate compared to propensity score models that ignored the multilevel structure. When the intra-class correlation was medium and the cluster-specific variation was properly accounted for, within-cluster matching and across-cluster matching provided similar results, but the within-cluster matching strategy offered the most accurate uncertainty estimate under the optimal matching method. When the intra-class correlation was low, the across-cluster matching strategy provided less biased treatment effect estimates. For the within-cluster matching strategy, the increase in the individual-level sample size offered better treatment effect and variation estimates. Prior in the propensity score model was shown to have little impact on the treatment effect estimation.

Results of the case study revealed the practical limitation of the within-cluster

matching strategy when there are too few subjects in one or both treatment conditions to facilitate close matches. Across-cluster matching can borrow information from subjects of other clusters when the within-cluster sample size is small and strength of treatment selection is not high. However, the assumption of all the cluster-level covariates being incorporated in the propensity score model for across-cluster matching is also hard to meet in practice. Violation of the strong ignorability assumption may still lead to biased causal effect estimates. Nevertheless, the Bayesian multilevel propensity score approach provided overall smaller retention effect compared to the unadjusted retention effect estimate, which implies that the proposed approach may effectively reduce the initial selection bias.

Overall, Bayesian random intercept and slope propensity score model for the within-cluster matching strategy is recommended when the within-cluster sample size is sufficiently large. When there is little evidence of omitted cluster-level covariates, Bayesian multilevel model with across-cluster matching may provide as good treatment effect and variation estimates. When within-cluster sample size is too small to facilitate close matches within each cluster, the across-cluster matching strategy can be an alternative when the cluster-level covariates are well controlled.

Note that the multilevel propensity score framework delineated in this dissertation can be applied to longitudinal observational studies with time-varying treatments but need extra assumption, i.e., sequential strong ignorability, and have some new challenges such as the presence of time-varying confounders that are outcomes of prior treatments but also predictors of later treatment assignments (Hong and Raudenbush, 2008). The methodology for longitudinal nonrandomized studies is beyond the scope of this dissertation and is warranted for future research.

Also, this dissertation is an early step of studying Bayesian multilevel propensity score approach for making causal inference in multilevel observational studies, which provided a practical Bayesian approach and studied its properties via different parametric propensity score models, sample sizes, priors, intra-class correlations, matching strategies and propensity score methods. Research can be conducted to further investigate the performance of Bayesian multilevel propensity score approach through non-parametric propensity score and outcome models,

with heterogeneous treatment effects across clusters and using different matching strategies such as a two-stage matching (Rickles, 2012), to advance in the causal inference for multilevel observational studies.

A R CODE FOR THE BAYESIAN MULTILEVEL PROPENSITY SCORE APPROACH IN THE CASE STUDY

```
## Note: A Bayesian multilevel PS model refers to a Bayesian random intercept
##       and slope propensity score model and a multilevel outcome model refers
##       to a random intercept outcome model here. The code can be easily modified for
##       other choices of models such as a Bayesian random intercept PS model.

## Load R packages (Need to install them if they were not installed before)
require(mice)
require(lme4)
require(MCMCpack)
require(gdata)
require(optmatch)
require(rjags) #need to install JAGS first

## Run the R functions needed for the analyses here.
## Functions are included at the bottom of this code.

## Read and code data
data1<-read.csv(file.choose(),header=TRUE)#Browse computer to select the analytic data set
data1=data1[-c(2,3)]

## Impute data
set.seed(2014)
imputed1<-mice(data1[,-c(14,17)],m=1)
newmod=glm.mids(C4RRSCAL~kretain+ t1arsgen+t2arsgen+
c2rscale+ p1learn+ t2rdgplo+ p1ageent+ female+p2tuitio+ b1klrn+ p1disabl+
a1prboys+ retention_rate+s2_id,data=imputed1)
temp=(newmod$analyses)
data_imputed1=temp[[1]]$data #Get 1 imputed data set for the individual-level data

set.seed(2014)
imputed2<-mice(data1[,c(1,14,17)],m=1)
newmod=glm.mids(c2rscale_mean~s2togthr+s2_id,data=imputed2)
temp=(newmod$analyses)
data_imputed2=temp[[1]]$data #Get 1 imputed data set for the school-level data
data_imputed=cbind(data_imputed1,data_imputed2)
data_imputed=data_imputed[,-c(15,16)]

y<-data_imputed$C4RRSCAL
x1<- data_imputed$c2rscale
```

```
x2<- data_imputed$t1arsgen
x3<- data_imputed$t2arsgen
x4<- data_imputed$p1learn
x5<- data_imputed$t2rdgplo
x6<- data_imputed$p1ageent
x7<- data_imputed$female
x8<- data_imputed$p2tuitio
x9<- data_imputed$p1disabl
x10<-data_imputed$a1prboys
x.11<-data_imputed$b1klrn
x12<- data_imputed$c2rscale_mean
x13<- data_imputed$s2togthr
retention<-data_imputed$kretain

## Dummy code the categorical variable
x13_2<-ifelse(x13==2,1,0)
x13_3<-ifelse(x13==3,1,0)
x13_4<-ifelse(x13==4,1,0)
x13_5<-ifelse(x13==5,1,0)


## Create a vector of consecutive school ID for each subject
data_imputed$s2_id<-as.factor(data_imputed$s2_id)
SchoolID=as.numeric(data_imputed$s2_id)

n.s=length(unique(SchoolID))
n.subj=length(y)
data_imputed_16<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x.11,x12,x13_2,
                            x13_3,x13_4,x13_5,SchoolID,retention,y)

## Bayesian unadjusted retention effect
set.seed(2014)
unadj<-MCMCregress(y~retention,data=data_imputed_16)
summary(unadj)# -19.52, 0.87

## Bayesian covariate-adjusted retention effect
set.seed(2014)
ancova<-MCMCregress(y~retention+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x.11+x12+
    x13_2+x13_3+x13_4+x13_5, data=data_imputed_16,burnin=5000)
summary(ancova) # -9.85, 0.66
```

```
#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
#::Bayesian Multilevel Propensity Score Model using JAGS::::::::::
#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

## Specify data
z<-matrix(0,n.s,3)
dat <- list(x1=x1,x2=x2,x3=x3,x4=x4,x5=x5,x6=x6,x7=x7,x8=x8,x9=x9,
            x10=x10,x.11=x.11,x12=x12,x13_2=x13_2,
            x13_3=x13_3,x13_4=x13_4,x13_5=x13_5,SchoolID=SchoolID,n.s=n.s,
            retention=retention,n.subj=n.subj,R1=diag(1,3,3),z=z)


# Specify model
modelstring = "
model {
for(i in 1:n.subj)
{
retention[i] ~ dbin (p[i], 1)
p[i] <- exp(lp[i])/(1+exp(lp[i]))
lp[i] <- a+b0[SchoolID[i],1] +b1*x1[i]+b2*x2[i]+b0[SchoolID[i],2]*x3[i]+
b4*x4[i]+b5*x5[i]+b6*x6[i]+b7*x7[i]+b8*x8[i]+b9*x9[i]+b10*x10[i]+
b0[SchoolID[i],3]*x.11[i]+b12*x12[i] + b13_2*x13_2[i]+
b13_3*x13_3[i] + b13_4*x13_4[i]+ b13_5*x13_5[i]}
for (j in 1:n.s) {
b0[j,1:3]~dmnorm(z[j,1:3],phi[1:3,1:3])
}

# Prior Specification
a~dnorm(0, 0.0001)
b1~dnorm(0, 0.0001)
b2~dnorm(0, 0.0001)
b4~dnorm(0, 0.0001)
b5~dnorm(0, 0.0001)
b6~dnorm(0,0.0001)
b7~dnorm(0, 0.0001)
b8~dnorm(0, 0.0001)
b9~dnorm(0, 0.0001)
b10~dnorm(0, 0.0001)
b12~dnorm(0, 0.0001)
b13_2~dnorm(0,0.0001)
b13_3~dnorm(0,0.0001)
b13_4~dnorm(0,0.0001)
b13_5~dnorm(0, 0.0001)
phi[1:3,1:3] ~ dwish(R1[1:3,1:3], 3)
```

```
}
"
writeLines(modelstring,con="model.bug")

## Run MCMC Chain
## Initialize Model
parameters = c("p","lp")
adaptSteps = 10000
burnInSteps = 10000
nChains =1
thinSteps = 100
nPerChain = 100000
Model1 = jags.model(file="model.bug",data=dat,n.chains=nChains, n.adapt=adaptSteps)

## Obtain the Posterior Samples of PS:
cat("Burning in the MCMC chain ...\n")
update(Model1, n.iter=burnInSteps)
cat("Sampling from the final MCMC chain ... \n")
codaSamples1 = coda.samples(Model1, variable.names=parameters,
n.iter=nPerChain, thin=thinSteps,seed=2014)
colnames(codaSamples1[[1]])

## Diagnostics and Plots
# Selected Trace plots and Density plots
plot(codaSamples1[[1]][,c(1536,1541,3047,3062)])

# Selected Autocorrelation plots
par(mfrow=c(2,2))
acf(codaSamples1[[1]][,1536],main="Child 0005023C")
acf(codaSamples1[[1]][,1541],main="Child 0007001C")
acf(codaSamples1[[1]][,3047],main="Child 3081020C")
acf(codaSamples1[[1]][,3062],main="Child 3113021C")

# Geweke diagnostic for the posterior propensity scores
geweke.diag(codaSamples1[[1]][,c(1536,1541,3047,3062)])
geweke<-geweke.diag(codaSamples1[[1]][,1532:3062])
sum(geweke$z<=1.96)/1531 #percentage of posterior PS that converged well
summary(geweke$z)

# Geweke Plot
geweke.plot(codaSamples1[[1]][,c(1536,1541,3047,3062)],frac1 = 0.1, frac2 = 0.5)

# Heidelberg-Welch Diagnostic
heidel<-heidel.diag(codaSamples1[[1]][,1532:3062])
```

```
#percentage of posterior PS that passed Heidelberg & Welch
(sum(na.omit(heidel[,2]==1)))/1531

## Posterior propensity scores
bps_RS<-t(codaSamples1[[1]][,(n.subj+1):(2*n.subj)])#Bayesian posterior PS
blps_RS<-t(codaSamples1[[1]][,1:n.subj])            #Logit of Bayesian posterior PS
n.posterior<-nrow(codaSamples1[[1]])                #Number of posterior draws of PS



#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
#::Conventional Outcome Model--The Approximate Bayesian Approach:::
#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

##==================Within-Cluster Matching=======================
## Intialization
G<-n.s #Number of clusters/schools
bmat_RS_sch<-cbind(rep(0,G),rep(0,G)) #School-specfic PS

data_imputed_16_sch<-rep(list(0),G)
for (g in 1:G){
  data_imputed_16_sch[[g]]<-subset(data_imputed_16,
  data_imputed_16$SchoolID==g)
}

## Optimal full matching
Bmat_RS<-cbind(rep(0,n.posterior),rep(0,n.posterior))
for (g in 1:G){
  for(i in 1:n.posterior)
  {
    Bmat_RS[i,]=optm(data_imputed_16_sch[[g]],
    data_imputed_16_sch[[g]]$y,
    bps_RS[data_imputed_16$SchoolID==g,i],
    data_imputed_16_sch[[g]]$retention)
}
  bmat_RS_sch[g,]=c(mean(Bmat_RS[,1]),sqrt(var(Bmat_RS[,1])+
  mean(Bmat_RS[,2]^2)))
}

## Weighted Mean trt effect and se estimate--weighted by inverse variance
(trt<-weighted.mean(x=bmat_RS_sch[,1],w=1/(bmat_RS_sch[,2]^2)))
(se<-sqrt(1/(sum(1/(bmat_RS_sch[,2]^2)))))
## 95% Interval-Assuming trt effect is normally distributed
c(trt-1.96*se,trt+1.96*se)
```

```
##==================Across-Cluster Matching=======================
## Optimal full matching
Bmat_RS<-cbind(rep(0,n.posterior),rep(0,n.posterior))#sinle-level outcome model
Bmat_RS_RI<-cbind(rep(0,n.posterior),rep(0,n.posterior))#multilevel outcome model

for(i in 1:n.posterior)
{
  Bmat_RS[i,]=optm(data_imputed_16,data_imputed_16$y,
  bps_RS[,i], data_imputed_16$retention)
  Bmat_RS_RI[i,]=optm_RI(data_imputed_16,
  data_imputed_16$y, bps_RS[,i], data_imputed_16$retention)
}

# Trt effect and SE
(bmat_RS =c(mean(Bmat_RS[,1]),sqrt(var(Bmat_RS[,1])+
mean(Bmat_RS[,2]^2))))
(bmat_RS_RI =c(mean(Bmat_RS_RI[,1]),sqrt(var(Bmat_RS_RI[,1])+
mean(Bmat_RS_RI[,2]^2))))

# 95% Intervals
(interval_RS_across<-c(mean(Bmat_RS[,1])-
1.96*sqrt(var(Bmat_RS[,1])+mean(Bmat_RS[,2]^2)),
mean(Bmat_RS[,1])+1.96*sqrt(var(Bmat_RS[,1])+mean(Bmat_RS[,2]^2))))

(interval_RS_RI_across<-c(mean(Bmat_RS_RI[,1])-
1.96*sqrt(var(Bmat_RS_RI[,1])+mean(Bmat_RS_RI[,2]^2)),
mean(Bmat_RS_RI[,1])+1.96*sqrt(var(Bmat_RS_RI[,1])+mean(Bmat_RS_RI[,2]^2))))




#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
#:::::::Bayesian Outcome Model--The Full Bayesian Approach:::::::::
#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::

##==================Within-Cluster Matching=======================
Bmat2_RS<-cbind(rep(0,n.posterior),rep(0,n.posterior))
bmat2_RS_sch<-cbind(rep(0,G),rep(0,G))

for (g in 1:G){
  for(i in 1:n.posterior)
  {
```

```
  Bmat2_RS[i,]=boptm(data_imputed_16_sch[[g]],
  data_imputed_16_sch[[g]]$y,
  bps_RS[data_imputed_16$SchoolID==g,i],
  data_imputed_16_sch[[g]]$retention)
  }
  bmat2_RS_sch[g,]=c(mean(Bmat2_RS[,1]),
  sqrt(var(Bmat2_RS[,1])+mean(Bmat2_RS[,2]^2)))
}

## Weighted Mean trt effect and se estimate--weighted by inverse variance
(btrt<-weighted.mean(x=bmat2_RS_sch[,1],w=1/(bmat2_RS_sch[,2]^2)))
(bse<-sqrt(1/(sum(1/(bmat2_RS_sch[,2]^2)))))
## 95% Interval
c(btrt-1.96*bse,btrt+1.96*bse)


##==================Across-Cluster Matching=======================
#Bayesian single-level outcome model
B2mat_RS<-cbind(rep(0,n.posterior),rep(0,n.posterior))

#Baysian multilevel outcome model
B2mat_RS_RI<-cbind(rep(0,n.posterior),rep(0,n.posterior))

for(i in 1:n.posterior)
{
  # Bayesian single-level model using MCMCregress
  B2mat_RS[i,]=boptm(data_imputed_16,
  data_imputed_16$y, bps_RS[,i], data_imputed_16$retention)
  #Bayesian multilevel model using rjags, fast and good convergence.
  data_imputed_16$pscore =bps_RS[,i]
  psdistance <- makedist(retention~1, data_imputed_16, scalardiffs, "pscore")
  fm=fullmatch(psdistance,data=data_imputed_16)
  strata=as.numeric(as.factor(fm))#obtain numeric label for each factor
  C=length(unique(strata))#number of stratas

  # Specify model
  modelstring = "
  model {
  for(i in 1:n.subj)
{
  y[i] ~ dnorm (mu[SchoolID[i]]+a*trt[i]+b[strata[i]], tausq_y)
}
  #Prior Specification
  for(j in 1:G)
```

```
{
  mu[j]~dnorm(alpha,tausq_mu)
}
  for (k in 1:C)
{
  b[k]~dnorm(0,0.001)
}
  a~dnorm(0, 0.001)
  tausq_y~dgamma(0.001,0.001)
  tausq_mu~dgamma(0.001,0.001)
  alpha~dnorm(0,0.0001)
  }
  "
  writeLines(modelstring,con="model.bug")

  # Run MCMC Chain
  parameters = c("a","mu")
  adaptSteps = 5000
  burnInSteps = 5000
  nChains =1
  thinSteps = 10
  nPerChain = 10000

  dat2=list(y=y, n.subj=n.subj, trt=retention, strata=strata,
  G=n.s, SchoolID=SchoolID,C=C)
  Model2 = jags.model(file="model.bug",data=dat2,n.chains=nChains,
  n.adapt=adaptSteps)
  cat("Burning in the MCMC chain ...\n")
  update(Model2, n.iter=burnInSteps)
  cat("Sampling from the final MCMC chain ... \n")
  codaSamples2 = coda.samples(Model2, variable.names=parameters,
                   n.iter=nPerChain, thin=thinSteps,seed=2014)
  effect<-codaSamples2[[1]][,1]
  B2mat_RS_RI[i,]=c(mean(effect),sd(effect))
}

## Trt effect and SE
(b2mat_RS =c(mean(B2mat_RS[,1]),sqrt(var(B2mat_RS[,1])+
mean(B2mat_RS[,2]^2))))
(b2mat_RS_RI =c(mean(B2mat_RS_RI[,1]),
 sqrt(var(B2mat_RS_RI[,1])+mean(B2mat_RS_RI[,2]^2))))

## 95% Intervals
(interval2_RI_across<-c(mean(B2mat_RS[,1])-
```

```
1.96*sqrt(var(B2mat_RS[,1])+mean(B2mat_RS[,2]^2)),
mean(B2mat_RS[,1])+1.96*sqrt(var(B2mat_RS[,1])+mean(B2mat_RS[,2]^2))))

(interval2_RI_RI_across<-c(mean(B2mat_RS_RI[,1])-
1.96*sqrt(var(B2mat_RS_RI[,1])+mean(B2mat_RS_RI[,2]^2)),
mean(B2mat_RS_RI[,1])+1.96*sqrt(var(B2mat_RS_RI[,1])+mean(B2mat_RS_RI[,2]^2))))

##======================== The END ========================



##================Functions that need to be run first======
# Functions for Optimal full matching that comes from the "optmatch" package.
# Attached here so that the code works regardless of which R version is used.
scalardiffs <- function(trtvar,data,scalarname) {
  sclr <- data[names(trtvar), scalarname]
  names(sclr) <- names(trtvar)
  abs(outer(sclr[trtvar],sclr[!trtvar], '-'))
}

makedist <- function(structure.fmla, data,
            fn=function(trtvar, dat, ...){
            matrix(0, sum(trtvar), sum(!trtvar),
            dimnames=list(names(trtvar)[trtvar],
            names(trtvar)[!trtvar]))},
                     ...)
{
  if (!attr(terms(structure.fmla), "response")>0)
    stop("structure.fmla must specify a treatment group variable")
  fn <- match.fun(fn)

  ### WHEN THIS FUNCTION IS WRAPPED TO, THIS IS HOW INFO ABOUT WHICH
  ### GENERATION PARENT FRAME structure.fmla IS TO BE EVALUATED IN IS PASSED
  pframe.generation <- 1
  if (!is.null(attr(structure.fmla, "generation.increment")))
    pframe.generation <- pframe.generation +
    attr(structure.fmla, "generation.increment")

  zpos <- attr(terms(structure.fmla), "response")
  vars <- eval(attr(terms(structure.fmla), "variables"), data,
               parent.frame(n=pframe.generation))
  zzz <- vars[[zpos]]
  if (!is.numeric(zzz) & !is.logical(zzz))
    stop("treatment variable (LHS of structure.fmla) must be numeric or logical")
```

```
if (any(is.na(zzz)))
  stop("NAs not allowed in treatment variable (LHS of structure.fmla)")
if (all(zzz>0))
  stop("there are no controls (LHS of structure.fmla >0)")
if (all(zzz<=0))
  stop("there are no treatment group members (LHS of structure.fmla <=0)")

zzz <- (zzz>0)
vars <- vars[-zpos]
names(zzz) <- row.names(data)
if (length(vars)>0)
{
  ss <- interaction(vars, drop=TRUE)
} else ss <- factor(zzz>=0, labels="m")
ans <- tapply(zzz, ss, FUN=fn,
              dat=data, ..., simplify=FALSE)
FUNchk <- unlist(lapply(ans,
                        function(x){!is.matrix(x) & !is.vector(x)}))
if (any(FUNchk)) { stop("fn should always return matrices")}

mdms <- split(zzz,ss)
NMFLG <- FALSE

for (ii in (1:length(ans)))
{
  dn1 <- names(mdms[[ii]])[mdms[[ii]]]
  dn2 <- names(mdms[[ii]])[!mdms[[ii]]]
  if (is.null(dim(ans[[ii]])))
  {
    if (length(dn1)>1 & length(dn2)>1)
    { stop("fn should always return matrices")}
    if (length(ans[[ii]])!=max(length(dn1), length(dn2)))
    { stop(paste("unuseable fn value for stratum", names(ans)[ii]))}

    if (is.null(names(ans[[ii]])))
    {
      ans[[ii]] <-  matrix(ans[[ii]], length(dn1), length(dn2),
                           dimnames=list(dn1,dn2))
    } else
    { if (length(dn1)>1)
    {
      ans[[ii]] <- matrix(ans[[ii]],length(dn1), 1,
                          dimnames=list(names(ans[[ii]]), dn2))
    } else {
```

```
        ans[[ii]] <- matrix(ans[[ii]], 1, length(dn2),
                            dimnames=list(dn1, names(ans[[ii]])))
      }
      }
    } else {
      if (!all(dim(ans[[ii]])==c(length(dn1), length(dn2))))
      { stop(paste("fn value has incorrect dimension at stratum",
                   names(ans)[ii])) }
      if (is.null(dimnames(ans[[ii]])))
      {
        dimnames(ans[[ii]]) <- list(dn1, dn2)
        NMFLG <- TRUE
      } else {
        if (!all(dn1%in%dimnames(ans[[ii]])[[1]]) |
              !all(dn2%in%dimnames(ans[[ii]])[[2]]) )
        { stop(paste(
          "dimnames of fn value don't match unit names in stratum",
          names(ans)[ii])) }
      }

    }
  }

  if (NMFLG){
warning("fn value not given dimnames;
assuming they are list(names(trtvar)[trtvar], names(trtvar)[!trtvar])")}
  attr(ans, 'row.names') <- names(zzz)
  class(ans) <- c('optmatch.dlist', 'list')
  ans
}

# Conventional single-level outcome model
optm=function(data,y,ps,trt)
{
  data$pscore = ps
  psdistance <- makedist(trt~1, data, scalardiffs, "pscore")
  fm=fullmatch(psdistance,data=data)   ## optimal full matching
  if(length(unique(fm))>1)
  {
    lmest=summary(lm(y~as.factor(trt)+as.factor(fm),data=data))$coef[2,c(1,2)]
  }
  else
  {
    lmest=summary(lm(y~as.factor(trt),data=data))$coef[2,c(1,2)]
```

```
  }
  return(c(lmest))
}


# Conventional random intercept outcome model
optm_RI=function(data,y,ps,trt,SchoolID)
{
  data$pscore = ps
  psdistance <- makedist(trt~1, data, scalardiffs, "pscore")
  fm=fullmatch(psdistance,data=data)
  results=summary(lmer(y~as.factor(trt)+
  as.factor(fm)+(1|SchoolID),data=data))
  return(c(results@coefs[2,1:2]))
}


# Bayesian single-level outcome model
boptm=function(data,y,ps,trt)
{
  data$pscore = ps
  psdistance <- makedist(trt~1, data, scalardiffs, "pscore")
  fm=fullmatch(psdistance,data=data)
  lmest=(MCMCregress(y~as.factor(trt)+as.factor(fm),
  data=data, mcmc = 1000))[,2]
  return(c(mean(lmest),sd(lmest)))
}


# Bayesian random intercept outcome model using MCMChregress, for reference only.
# In the code, rjags was used instead for Bayesian random intercept outcome model
# for faster computation and better convergence.
boptm_RI=function(data,y,ps,trt,S_ID)
{
  data$pscore = ps
  psdistance <- makedist(trt~1, data, scalardiffs, "pscore")
  fm=fullmatch(psdistance,data=data)
  lmest=(MCMChregress(fixed=y~as.factor(trt)+as.factor(fm),
    random=~1,group="S_ID",data=data,
    burnin=5000, mcmc=10000, thin=10,
    mubeta=priormean,Vbeta=priorvar,r=1, R=36))$mcmc[,2]
  return(c(mean(lmest),sd(lmest)))
}
```

REFERENCES

Abadie, A., and G. W. Imbens. 2009. Matching on the estimated propensity score. *NBER working paper 15301*.

An, W. H. 2010. Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology* 40:151–189.

Arpino, B., and F. Mealli. 2011. The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis* 55: 1770–1780.

Barnard, J., C. Frangakis, J. Hill, and D. B. Rubin. 2003. A principal stratification approach to broken randomized experiments: A case study of vouchers in New York City (with discussion and rejoinder). *Journal of the American Statistical Association* 98:299–323.

Bates, D., M. Maechler, and B. Bolker. 2013. *lme4: Linear mixed-effects models using s4 classes*.

Bates, D., and D. Sarkar. 2007. *lme4: Linear mixed-effects models using s4 classes*.

Chen, J., and D. Kaplan. 2014. Covariate balance in bayesian propensity score approaches for observational studies. *Journal of Research on Educational Effectiveness*.

Chib, S., and E. Greenberg. 2010. Bayesian matching for causal inference. Unpublished manuscript.

Cochran, W. G. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24:205–213.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences, 2nd edition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

De Finetti, B. 1974. *Theory of probability, vols. 1 and 2*. New York: John Wiley and Sons.

Dedrick, R. F., J. M. Ferron, M. R. Hess, K. Y. Hogarty, J. D. Kromrey, T. R. Lang, J. D. Niles, and R. S. Lee. 2009. Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research* 79:69–102.

Dehejia, R. H., and S. Wahba. 1999. Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 94:1053–1062.

———. 2002. Propensity score-matching methods for non-experimental causal studies. *The Review of Economics and Statistics* 84:151–161.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38.

Dempster, A. P., D. B. Rubin, and R. K. Tsutakawa. 1981. Estimation in covariance components model. *Journal of the American Statistical Association* 76:341–353.

Draper, D. 1995. Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics* 20:115–147.

Efron, B., and C. Morris. 1975. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* 74:311–319.

Elston, R. C., and J. E. Grizzle. 1962. Estimation of time-response curves and their confidence bands. *Biometrics* 18:148–159.

Foster, E. M. 2003. Propensity score matching an illustrative analysis of dose response. *Medical Care* 41:1183–1192.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.

Gelman, A., Y-S. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, and T. Zheng. 2011. *Arm: Data analysis using regression and multilevel/hierarchical models*. R package version 1.4-13.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics*, ed. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, 4th ed. Oxford, U.K.: Oxford University Press.

Gilks, W. R., and D. J Spiegelhalter. 1996. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Goldstein, H. 1986. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73:43–56.

Griswold, M. E., A. R. Localio, and C. Mulrow. 2010. Propensity socre adjustment with multilevel data: Setting your sites on decreasing slection bias. *Annals of Internal Medicine* 152:393–395.

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

Guo, S., R. P. Barth, and C Gibbons. 2006. Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review* 28:357–383.

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., and S. O. Klopfer. 2006. Optimal full matching and related designs via network flow. *Journal of Computational and Graphical Statistics* 15:609–627.

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

Hartung, J., G. Knapp, and B. K. Sinha. 2008. *Statistical meta-analysis with applications*. John Wiley & Sons.

Heckman, J. J. 2005. The scientific model of causality. In *Sociological methodology*, ed. Ross. M. Stolzenberg, vol. 35, 1–97. Boston: Blackwell Publishing.

Heidelberger, P., and P. D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* 31:1109–1144.

Hirano, K., and G. W Imbens. 2001. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* 2:259–278.

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

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

———. 2006. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association* 101:901–910.

———. 2008. Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics* 33:333–362.

Hong, G., and B. Yu. 2007. Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis* 29:239–261.

———. 2008. Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology* 44:407–421.

Horvitz, D. G., and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47: 663–685.

Hoshino, T. 2008. A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis* 52:1413–1429.

Hox, J. J. 2010. *Multilevel Analysis: Techniques and Applications, 2nd edition*. New York, NY: Routledge.

Kang, J., and J. L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22:523–539.

Kaplan, D. 2014. *Bayesian Statistics for the Social Sciences*. New York: Guilford.

Kaplan, D., and J. Chen. 2012. A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika* 77:581–609.

Kelcey, B. 2009. Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings. `http://hdl.handle.net/2027.42/63716`. Doctoral dissertation, University of Michigan.

Kim, J., and M. Seltzer. 2007. Causal inference in multilevel settings in which selection processes vary across schools. Tech. Rep., UCLA.

Kish, L. 1965. *Survey sampling*. New York, NY: Wiley.

Kurth, T., A. M. Walker, R. J. Glynn, K. A. Chan, M. J. Gaziano, K. Berger, and J. M Robins. 2006. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiololy* 163:262–270.

Lee, B. K., J. Lessler, and E. A. Stuart. 2009. Improving propensity score weighting using machine learning. *Statistics in Medicine* 29:337–346.

———. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine* 29:337–346.

Leow, C., S. Marcus, E. Zanutto, and R Boruch. 2004. Effects of advanced course-taking on math and science achievement: Addressing selection bias using propensity scores. *American Journal of Evaluation* 25:461–478.

Li, F., A. M. Zaslavsky, and M. B. Landrum. 2013. Propensity score weighting with multilevel data. Unpublished manuscript.

Lindley, D. V., and A. F. M. Smith. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B* 34:1–41.

Lindstrom, M., and D. Bates. 1988. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 83:1014–1022.

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

Longford, N. T. 1987. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* 74:817–827.

———. 1990. *VARCL: Software for variance component analysis of data with nested random effects (maximum likelihood)*. Princeton, NJ: Educational Testing Service.

———. 1993. *Random coefficient models*. Oxford, UK: Clarendon Press.

Lunceford, J.K., and M Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23:2937–2960.

Lunn, D., N. Best, D. Spiegelhalter, G. Graham, and B. Neuenschwander. 2009. Combining mcmc with "sequential" PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics* 36:19–38.

Martin, A. D., K. M. Quinn, and J. H. Park. 2011. MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software* 42(9):22–44.

McCaffrey, D. F., G. Ridgeway, and A. R Morral. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 9:403–425.

McCandless, L. C., I. J. Douglas, S. J. Evans, and L. Smeeth. 2010. Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics* 6:Article 16.

McCandless, L. C., P. Gustafson, and P. C. Austin. 2009. Bayesian propensity score analysis for observational data. *Statistics in Medicine* 28:94–112.

McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models, 2nd edition*. London: Chapman & Hall.

NCES. 2001. Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual (tech. rep. no. nces 2001-029). Tech. Rep., U.S. Department of Education.

Neyman, J. S. 1923. Statistical problems in agriculture experiments. *Journal of the Royal Statistical Society, Series B* 2:107–180.

Patterson, H. D., and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. *Biometrika* 58:545–554.

Plummer, M. 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling. Vienna, Austria: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).

———. 2014. *rjags: Bayesian graphical models using mcmc*. R package version 3-13.

Pruzek, R. M. 2011. Introduction to the special issue on propensity score methods in behavioral research. *Multivariate Behavioral research* 46:389–398.

R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, C. R. 1972. Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association* 67:112–115.

Rasbash, J., W. J. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, and A. Apple. 2000. *A user's guide to MLwiN*. London: Multilevel Models Project, University of London.

Raudenbush, S. W., and A. S. Bryk. 1986. A hierarchical model for studying school effects. *Sociology of Education* 59:1–17.

———. 2002. *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Raudenbush, S. W., A. S. Bryk, Y. F. Cheong, and R. T. Congdon. 2004. *HLM 6: Hierarchical linear and nonlinear modeling*. Scientific Software International Inc.

Rickles, J. H. 2012. Using a two-stage propensity score matching strategy and multilevel modeling to estimate treatment effects in a multisite observational study. `http://gradworks.umi.com/35/11/3511076.html`. Doctoral dissertation, University of California, Los Angeles.

Robinson, W. S. 1950. Ecological correlations and the behavioral of individuals. *American Sociological Review* 15:351–357.

Rose, S. 2012. Third grade reading policies. Tech. Rep., Denver, CO: Education Commission of the States.

Rosenbaum, P. R. 1987. Model-based direct adjustment. *Journal of the American Statistical Association* 82:387–394.

———. 1989. Optimal matching for observational studies. *Journal of the American Statistical Association* 84:1024–1032.

———. 2002. *Observational studies (2nd ed.)*. New York, NY: Springer-Verlag.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Rosenbaum, P. R., and D. B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79:516–524.

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

———. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74:318–328.

Rubin, D. B. 1985. The use of propensity scores in applied Bayesian inference. *Bayesian Statistics* 2:463–472.

Rubin, D. B. 1986. Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers . *Journal of the American Statistical Association* 81:961–962.

SAS Institute, Inc. 2008. *SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schafer, J. L., and J. Kang. 2008. Average causal effects from non-randomized studies: A practical guide and simulated example. *Psychological Methods* 13:279–313.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance components*. Hoboken, NJ: John Wiley & Sons.

Shadish, W. R., M. H. Clark, and P. M. Steiner. 2008. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association* 103: 1334–1356.

Snijders, T. A. B., and R. Bosker. 2012. *Multilevel analysis: An introduction to basic and advanced multilevel modeling, 2nd edition*. Thousand Oaks, CA: Sage.

Spiegelhalter, D. J., A. Thomas, and N. G. Best. 1999. *Winbugs version 1.2 user manual*. MRC Biostatistics Unit.

Steiner, P. M., and D. Cook. 2013. Matching and propensity scores. In *Oxford handbook of quantitative methods*, ed. T. D. Little, vol. 1. New York, NY: Oxford University Press.

Steiner, P. M., J. Kim, and F. Thoemmes. 2013. Matching strategies for observational multilevel data. In *JSM Proceedings*, 5020–5032. Alexandria, VA: American Statistical Association.

Sutton, J. R. 2012. Imprisonment and Opportunity Structures: A Bayesian Hierarchical Analysis. *European Sociological Review* 28:12–27.

Swanson, J. M., S. P. Hinshaw, E. Arnold, R. D Gibbons, S. Marcus, K Hur, and P. S Jensen. 2007. Secondary evaluations of mta 36-month outcomes: Propensity score and growth mixture model analyses. *Journal of the American Academy of Child & Adolescent Psychiatry* 46:1003–1014.

Thoemmes, F. J., and E. S. Kim. 2011. A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research* 46:90–118.

Thoemmes, F. J., and S. G. West. 2011. The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research* 46:514–543.

Van Buuren, S., and K. Groothuis-Oudshoorn. 2011. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45:1–67.

Wu, C. F. J. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics* 11:95–103.

Wu, W., S. G. West, and J. N. Hughes. 2010. Effect of grade retention in first grade on psychosocial outcomes. *Journal of Educcational Psychology* 102(10):135–152.

Xu, Z., and J. D. Kalbfleisch. 2010. Propensity score matching in randomized clinical trials. *Biometrics* 66:813–823.

Yanovitzkya, I., E. Zanuttob, and R Hornikb. 2005. Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning* 28:209–220.

Zigler, C. M., K. Watts, R. W. Yeh, Y. Wang, B. A. Coull, and F. Dominici. 2013. Model feedback in bayesian propensity score estimation. *Biometrics* 69:263–273.

Zill, N., L. S. Loomis, and J. West. 1997. The elementary school performance and adjustment of children who enter kindergarten late or repeat kindergarten: Findings from national surveys (statistical analysis report nces 98-097). Tech. Rep., Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.