

Causal Inference for Multi-level Observational Data with Application to Kindergarten Retention Study

Guanglei Hong and Stephen W. Raudenbush

School of Education, University of Michigan, Ann Arbor, Michigan 48109-1259

KEY WORDS: Potential Outcomes; Exchangeability; Propensity Scores; Multi-Site Trial; Cluster Randomized Trial

Challenges of Multi-Level Data to Rubin's Causal Framework

Rubin's (1978) potential-outcomes causal framework laid the foundation for conceptualizing causal problems and developing statistical solutions. For simplicity, he presented the framework under the stable unit treatment value assumption (SUTVA). It assumes that there is a single value of each potential outcome associated with each treatment for each experimental unit, regardless of how the treatments are assigned and what treatments are received by other experimental units. At the core of this causal inference framework is the ignorability assumption. In particular, the treatment assignment mechanism is considered to be strongly ignorable if, given the observed covariates and outcomes, the probability of treatment assignment is independent of all the unobserved covariates and potential outcomes.

Rosenbaum and Rubin (1983) used the propensity score to represent the treatment assignment mechanism. When the strong ignorability assumption holds, the propensity score is the conditional probability of treatment assignment given the observed covariates. In non-experimental studies, the propensity score is to be estimated from the observed data. Rosenbaum and Rubin (1983) proved that, when we hold the propensity score constant, there should be no difference in the distribution of any pretreatment covariate between two treatment groups. In the past two decades, researchers have proposed a variety of propensity score-based approaches to causal inference for observational data under strong ignorability and SUTVA. These include propensity score matching (Rosenbaum, 1989, 2002; Rubin & Thomas, 1992, 1996), stratification (Rosenbaum, 1991; Rosenbaum & Rubin, 1984), covariance adjustment (Rosenbaum & Rubin, 1983; Rubin & Thomas, 2000), and inverse-probability-of-treatment weighting (IPTW) (Robins, 1997, 2000).

Most researchers simply assume that the SUTVA is true without further theoretical or empirical scrutiny. In general, this assumption is hardly tenable when treatment assignment occurs in a multi-level setting. This is partly due to the sharing of and competition for resources within and between organizations (Cook, 1977; Emerson, 1962; Pfeffer & Salancik, 1978), and partly due to the agent effects in

treatment delivery (Lipsky, 1980; Manski & Garfinkel, 1992). Conceptually there is a distinct set of potential outcomes associated with each treatment for each unit corresponding to all the possible group composition, agent allocation, and treatment allocation. The multiplicity of potential outcomes in a multi-level setting is an important theoretical issue in causal inference. Also yet to be explored is the applicability of the propensity score-based causal inference methods for multi-level data.

The primary objectives of this study are to extend current theory in statistical science about causal inference to encompass multi-level data and to investigate the implications of this theoretical extension for the propensity score-based causal inference techniques. For illustration, we apply the extended framework and the statistical techniques to an empirical study of the causal effects of kindergarten retention on student learning using multi-level observational data. We explore answers to the following theoretical questions: (1) How shall we define the causal effect in a multi-level setting when the SUTVA does not hold? (2) What are the treatment assignment mechanisms in multi-level experimental designs? Correspondingly, what are the potential sources of bias in multi-level non-experimental data? (3) What are the possible consequences of mistaking one multi-level design for another? (4) How shall we apply the propensity score-based methods to the causal inference for multi-level observational data?

Exchangeability Assumption

In this section, we relax the SUTVA, and invoke a comparatively simple assumption that is benign for defining and estimating causal effects of treatments in a multi-level setting.

For unit i that is randomly drawn from a population, suppose that there are two treatments, $Z_i = 1$ if unit i is assigned to the experimental group and $Z_i = 0$ if i is assigned to the control group. Under the SUTVA, every unit has only two potential outcomes corresponding to these two treatments. The potential outcome models can be written as follows.

$$\begin{aligned} Y_i^{(Z_i=0)} &= \mu + \varepsilon_i^{(Z_i=0)}, \\ Y_i^{(Z_i=1)} &= \mu + \delta + \varepsilon_i^{(Z_i=1)}. \end{aligned} \quad (1)$$

Here μ is the population average potential outcome associated with the treatment $Z = 0$, δ is the population average treatment effect, $\varepsilon_i^{(0)}$ and $\varepsilon_i^{(1)}$ are unit-specific random effects that are assumed to have

a mean of zero and a constant variance of σ_Y^2 . We usually assume a zero correlation of the outcome observations between every two units.

When the experimental units are clustered to receive individual-level or cluster-level treatments, the potential outcomes for an experimental unit will at least depend on which units are assigned to the same cluster and who delivers the treatment.¹ The potential outcomes for an experimental unit in a certain cluster may no longer be constant if we change the cluster assignment and agent assignment. Hence, when we relax the SUTVA, we face the problem of having a vast number of potential outcomes for each experimental unit.

Assuming additivity, we can specify each potential outcome for each unit as a sum of the expected outcome of the population given the treatment, the unit-specific random effect, and the cluster-specific random effect. Following Lindley and Smith (Lindley, 1972; Lindley & Smith, 1972), we view the distribution of cluster-specific random effects as exchangeable when no prior knowledge suggests otherwise. In other words, the observed group composition and agent allocation, along with their consequent effects on the potential outcomes, are viewed as random events that are exchangeable with all other possible events that have the same prior distribution of occurrence. This is similar to having “a random sample from some distribution, with a (prior) distribution over the sampled distribution” (Lindley, 1972, p.39). Typically, the prior distribution that we impose on the cluster-level random effects has a mean of zero and a constant variance. We further assume that there is no interference between clusters and that, for a given cluster, the observations within that cluster are independent. The population average causal effect is now averaged over all the exchangeable cluster-level random effects. As we will illustrate in the next section, the effects of both interference between units within clusters and treatment enactment variation across clusters have been incorporated into the cluster-specific treatment effect that has an expected mean of zero.

Randomized Experiments for Multi-Level Treatments

In this section we describe three typical types of multi-level experimental designs—the multi-site randomized design, cluster randomized design, and joint multi-level randomized design.² For each of these multi-level designs, we develop the potential-outcome models, and define the causal effects under the exchangeability assumption. Then we briefly discuss model-based methods for estimating the causal effects in each design.

A *multi-site randomized design* replicates a completely randomized design or a randomized block design over multiple experimental sites. Within each site, individuals or groups of individuals are assigned at random to different treatments (Raudenbush & Liu, 2000). For simplicity, we consider two alternative treatments, $Z = 1$ and $Z = 0$, to be assigned to individual units. Let P be the probability for a unit to be assigned to the experiment within a site. We use $Y_{ij}^{(Z_{ij}=0)}$ and $Y_{ij}^{(Z_{ij}=1)}$ to denote the potential outcomes corresponding to these two treatments for individual i in site j . Assuming additivity, the potential-outcome models can be written as follows.

$$\begin{aligned}
 Y_{ij}^{(Z_{ij}=0)} &= \gamma + u_{0j} + e_{ij}^{(Z_{ij}=0)}; \\
 Y_{ij}^{(Z_{ij}=1)} &= \gamma + u_{0j} + \delta + u_{1j} + e_{ij}^{(Z_{ij}=1)}; \\
 \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right); \\
 Z_{ij} e_{ij}^{(Z_{ij}=1)} + (1 - Z_{ij}) e_{ij}^{(Z_{ij}=0)} &\sim N(0, \sigma^2). \tag{2}
 \end{aligned}$$

As we defined before, γ is the population average potential outcome associated with the treatment $Z = 0$. However, δ is to be defined as the population average within-site treatment effect. Here u_{0j} is the site-specific random effect that does not depend on the treatments; u_{1j} is the site-specific increment to the treatment effect, which incorporates the effects of interference between units, agent effects, and all other site-level contextual effects that interact with the treatment effect. The individual-specific random effects are $e_{ij}^{(Z_{ij}=0)}$ and $e_{ij}^{(Z_{ij}=1)}$. Hence, the treatment effect for individual i at site j is the sum of the population average within-site treatment effect, δ , the site-specific increment to the treatment effect, u_{1j} , and the individual-specific random increment to the treatment effect, $e_{ij}^{(Z_{ij}=1)} - e_{ij}^{(Z_{ij}=0)}$.

$$\delta_{ij} = Y_{ij}^{(Z_{ij}=1)} - Y_{ij}^{(Z_{ij}=0)} = \delta + u_{1j} + [e_{ij}^{(Z_{ij}=0)} - e_{ij}^{(Z_{ij}=1)}]. \tag{3}$$

Note that the SUTVA holds within each site. Hence, the mean difference in the observed outcome between the experimental and the control units in site j is an unbiased estimator of the treatment effect in this site, $\hat{\delta}_j = \hat{\delta} + \hat{u}_{1j} = \bar{Y}_{.j}^{(Z=1)} - \bar{Y}_{.j}^{(Z=0)}$. However, if the experiment is replicated in more than two sites and if exchangeability can be assumed between regression equations for multiple sites, the above least-squares estimate is inadmissible (Lindley & Smith, 1972). A Bayes or empirical Bayes posterior estimate is almost certain to have a smaller expected sum of squared errors (Raudenbush, 1988). Since the treatment assignment is ignorable in a multi-site experimental design, we can obtain a consistent estimator of the treatment effect δ through two-level hierarchical linear modeling. The combined model is basically a two-way ANOVA mixed-effects model that contains an interaction between the fixed treatment effect and

the random site effect. The exchangeability assumption is assumed in this model-based approach. We use Y_{ij} to denote the observed outcome of unit i in site j .

$$Y_{ij} = \gamma_{00} + \delta Z_{ij} + u_{0j} + u_{1j} Z_{ij} + e_{ij}. \quad (4)$$

To test the null hypothesis of no treatment effect, that is, $\delta = 0$, we assess the treatment effect estimate against the estimate of its standard error through a t test. Multi-site trials also enable estimation of variation in treatment effect across sites. Such variation can possibly be attributed to treatment enactment variation or moderating effects of site characteristics or both. We test the null hypothesis of no variation in treatment effect, that is, $\tau_{11} = 0$, through a chi-square test.

In *cluster-randomized designs*, organizations are the experimental units that are assigned at random to different treatments. Individuals in the same cluster are observation units that provide repeated measures of the organization's overall response to the treatment received (Murray, 1998). For simplicity, let Q be the probability for a cluster to be assigned to the experiment. Then all the individuals in this cluster have the same probability, Q , of being assigned to the experiment. It is important to note that cluster-level randomization ensures that the treatments received by individual units within clusters are independent of their potential outcomes. Suppose that $D = 1$ and $D = 0$ are two alternative treatments to be assigned to clusters. The potential outcome models for individual i in cluster j can be written as follows.

$$\begin{aligned} Y_{ij}^{(Dj=0)} &= \gamma_{00} + u_j^{(Dj=0)} + e_{ij}^{(Dj=0)}, \\ Y_{ij}^{(Dj=1)} &= \gamma_{00} + \delta + u_j^{(Dj=1)} + e_{ij}^{(Dj=1)}. \end{aligned} \quad (5)$$

where $Y_{ij}^{(Dj=0)}$ is the potential outcome of individual i in cluster j if this cluster is assigned to the control treatment; $Y_{ij}^{(Dj=1)}$ is the potential outcome for the same individual in the same cluster that is alternatively assigned to the experimental treatment; γ_{00} is the population average potential outcome associated with the control condition $D = 0$; δ is the population average between-cluster treatment effect; $u_j^{(Dj=0)}$ is the cluster-specific increment to the average potential outcome if cluster j is assigned to the control treatment; $u_j^{(Dj=1)}$ is the cluster-specific increment if cluster j is assigned to the experimental treatment; $e_{ij}^{(Dj=0)}$ and $e_{ij}^{(Dj=1)}$ are random effects associated with individual i in cluster j corresponding to the control condition and the experimental condition, respectively. Hence, the treatment effect for individual i in cluster j is a sum of the population average between-cluster treatment effect, δ , the cluster-specific increment to the treatment effect, $u_j^{(Dj=1)} - u_j^{(Dj=0)}$, and the individual-specific increment to the treatment effect, $e_{ij}^{(Dj=1)} - e_{ij}^{(Dj=0)}$.

$$\delta_{ij} = \delta + [u_j^{(Dj=1)} - u_j^{(Dj=0)}] + [e_{ij}^{(Dj=1)} - e_{ij}^{(Dj=0)}];$$

$$\begin{aligned} D_j u_j^{(Dj=1)} + (1 - D_j) u_j^{(Dj=0)} &\sim N(0, \tau); \\ D_j e_{ij}^{(Dj=1)} + (1 - D_j) e_{ij}^{(Dj=0)} &\sim N(0, \sigma^2). \end{aligned} \quad (6)$$

Because the cluster-level treatment assignment is ignorable, we obtain a consistent estimate of the between-site treatment effect δ through two-level hierarchical linear modeling. The combined model is a two-way ANOVA mixed model.

$$Y_{ij} = \gamma_{00} + \delta D + u_j + e_{ij}. \quad (7)$$

In reality, often there are multiple sets of treatments conducted simultaneously at multiple levels. Despite the fact that *joint multi-level designs* often most closely resemble studies of social experiments in natural settings, this category of designs has not yet received much systematic theoretical examination in the literature. A simple example is a two-step experimental design in which clusters are assigned at random to the cluster-level policies, and individuals within a cluster are assigned at random to the individual-level treatments. The probability of individual-level treatment Z_{ij} for individual i in cluster j may depend on cluster-level policy D_j . Suppose that Z and D are both binary. Also for simplicity, suppose that the probability of assigning cluster j to the experimental condition is a constant, $\Pr(D_j = 1) = Q$. For individual i in cluster j that has been assigned to the cluster-level control condition, the probability of being assigned to the individual-level experiment within this cluster is $\Pr(Z_{ij} = 1 \mid D_j = 0) = q_0$. If cluster j is assigned to the experimental condition instead, the probability for individual i in this cluster to be assigned to the individual-level experiment is $\Pr(Z_{ij} = 1 \mid D_j = 1) = q_1$.

Because Z and D are both binary, we have four potential outcomes for individual i in cluster j .

$$\begin{aligned} Y_{ij}^{(Zij=0,Dj=0)} &= \gamma_{00} + u_{0j}^{(Dj=0)} + e_{ij}^{(Zij=0,Dj=0)}, \\ Y_{ij}^{(Zij=1,Dj=0)} &= \gamma_{00} + \delta_Z + u_{0j}^{(Dj=0)} + u_{1j}^{(Dj=0)} + e_{ij}^{(Zij=1,Dj=0)}, \\ Y_{ij}^{(Zij=0,Dj=1)} &= \gamma_{00} + \delta_D + u_{0j}^{(Dj=1)} + e_{ij}^{(Zij=0,Dj=1)}, \\ Y_{ij}^{(Zij=1,Dj=1)} &= \gamma_{00} + \delta_D + \delta_Z + \delta_{DZ} + u_{0j}^{(Dj=1)} + u_{1j}^{(Dj=1)} \\ &\quad + e_{ij}^{(Zij=1,Dj=1)}. \end{aligned} \quad (8)$$

Assuming exchangeability, we specify the variances and covariance of the random effects as follows.

$$\begin{aligned} D_j u_{0j}^{(Dj=1)} + (1 - D_j) u_{0j}^{(Dj=0)} &\sim N(0, \tau_{00}); \\ D_j u_{1j}^{(Dj=1)} + (1 - D_j) u_{1j}^{(Dj=0)} &\sim N(0, \tau_{11}); \\ \text{Cov}(u_{0j}, u_{1j}) &= \tau_{01}; \\ (1 - Z_{ij})(1 - D_j) e_{ij}^{(Zij=0,Dj=0)} + Z_{ij}(1 - D_j) e_{ij}^{(Zij=1,Dj=0)} \\ + (1 - Z_{ij}) D_j e_{ij}^{(Zij=0,Dj=1)} + Z_{ij} D_j e_{ij}^{(Zij=1,Dj=1)} &\sim N(0, \sigma^2). \end{aligned} \quad (9)$$

Here γ_{00} is the population average outcome associated with a combination of the cluster-level control condition $D = 0$ and the individual-level control treatment $Z = 0$; u_{0j} is the cluster-specific increment to γ_{00} ; and u_{1j} is the cluster-specific increment to the within-cluster treatment effect δ_Z . It is possible to model the heterogeneity of variance in each of these cluster-specific increments. For example, if we have

theoretical reasons to believe that the cluster-level treatment D_j may have impact on the variance of the potential outcomes for individuals at site j , we can specify that the respective variances of $u_{0j}^{(D_j=1)}$ and $u_{0j}^{(D_j=0)}$ are τ_{00} and τ_{00}' , and that the respective variances of $u_{1j}^{(D_j=1)}$ and $u_{1j}^{(D_j=0)}$ are τ_{11} and τ_{11}' . Nonetheless, constant variance is a benign assumption in most cases.

Note that, for individual i in cluster j , the within-cluster treatment effect may depend on the treatment condition at the cluster level.

$$\begin{aligned} \delta_{Zij}^{(D_j=0)} &= Y_{ij}^{(Z_{ij}=1, D_j=0)} - Y_{ij}^{(Z_{ij}=0, D_j=0)} = \delta_Z + u_{1j}^{(D_j=0)} \\ &\quad + (e_{ij}^{(Z_{ij}=1, D_j=0)} - e_{ij}^{(Z_{ij}=0, D_j=0)}); \\ \delta_{Zij}^{(D_j=1)} &= Y_{ij}^{(Z_{ij}=1, D_j=1)} - Y_{ij}^{(Z_{ij}=0, D_j=1)} = \delta_Z + \delta_{DZ} + u_{1j}^{(D_j=1)} \\ &\quad + (e_{ij}^{(Z_{ij}=1, D_j=1)} - e_{ij}^{(Z_{ij}=0, D_j=1)}). \end{aligned} \quad (10)$$

Here δ_Z is the population average within-cluster treatment effect under the control condition $D = 0$, whereas $\delta_Z + \delta_{DZ}$ is the population average within-cluster treatment effect under the experimental condition $D = 1$. Conversely, the cluster-level policy effect on individual i in cluster j may depend on the specific treatment that individual i is assigned to.

$$\begin{aligned} \delta_{Dij}^{(Z_{ij}=0)} &= Y_{ij}^{(Z_{ij}=0, D_j=1)} - Y_{ij}^{(Z_{ij}=0, D_j=0)} = \delta_D + (u_{0j}^{(D_j=1)} - u_{0j}^{(D_j=0)}) \\ &\quad + (e_{ij}^{(Z_{ij}=0, D_j=1)} - e_{ij}^{(Z_{ij}=0, D_j=0)}); \\ \delta_{Dij}^{(Z_{ij}=1)} &= Y_{ij}^{(Z_{ij}=1, D_j=1)} - Y_{ij}^{(Z_{ij}=1, D_j=0)} = \delta_D + \delta_{DZ} + (u_{0j}^{(D_j=1)} - u_{0j}^{(D_j=0)}) \\ &\quad + (u_{1j}^{(D_j=1)} - u_{1j}^{(D_j=0)}) + (e_{ij}^{(Z_{ij}=1, D_j=1)} - e_{ij}^{(Z_{ij}=1, D_j=0)}). \end{aligned} \quad (11)$$

Here δ_D is the population average between-cluster treatment effect for individuals who are assigned to the control condition $Z = 0$, and $\delta_D + \delta_{DZ}$ is the population average between-cluster treatment effect for those in the experimental condition $Z = 1$. The treatment effects δ_Z , δ_D , and δ_{DZ} can be estimated through a mixed-effects model as follows. As before, Y_{ij} is the observed outcome of unit i in cluster j .

$$Y_{ij} = \gamma_{00} + \delta_Z Z_{ij} + \delta_D D_j + \delta_{DZ} Z_{ij} D_j + u_{0j} + u_{1j} Z_{ij} + e_{ij}. \quad (12)$$

Possible Consequences of Mistaking Multi-Level Designs

What would be the consequences if data analysts, without full knowledge about the initial design, mistake a joint multi-level design for a multi-site design or a cluster design, or even for a completely randomized design? Would they still obtain an appropriate estimate of the individual-level treatment effect if they simply ignore the cluster-level treatment? Or vice versa? Answers to these questions will have important implications, especially for analyses of non-experimental data where the designs are rarely known a priori. We derive the results as follows. We first show that, when the cluster-level treatments ($Q \neq 0$) are ignored in a joint multi-level randomized design, unless under special conditions, the expected difference between the two potential outcomes corresponding to the two individual-level treatments, $Z = 1$ and $Z = 0$, is likely to be an

inappropriate estimand for the individual-level treatment effect, δ_Z .

$$\begin{aligned} E[Y_{ij}^{(Z_{ij}=1)} - Y_{ij}^{(Z_{ij}=0)}] &= \delta_Z \\ &+ \delta_D \frac{(q_1 - q_0)Q(1-Q)}{[q_1Q + q_0(1-Q)][(1-q_1Q) - q_0(1-Q)]} \\ &+ \delta_{DZ} \frac{q_1Q}{q_1Q + q_0(1-Q)}. \end{aligned} \quad (13)$$

$E[Y_{ij}^{(Z_{ij}=1)} - Y_{ij}^{(Z_{ij}=0)}]$ is equal to δ_Z only if one of the following two conditions is true: (a) $q_0 = q_1 = q$, and $\delta_{DZ} = 0$, that is, neither an individual's probability of receiving the individual-level treatment nor the effect of the individual-level treatment depends on the cluster-level treatment; or (b) $\delta_D = \delta_{DZ} = 0$, that is, the cluster-level treatment has no effect whatsoever. When neither of these two conditions is satisfied, the expected difference between the two potential outcomes, $Y_{ij}^{(Z_{ij}=1)}$ and $Y_{ij}^{(Z_{ij}=0)}$, for individual i in cluster j is a sum of the individual-level treatment effect, δ_Z , a proportion of the cluster-level treatment effect, δ_D , and a proportion of the interaction effect between the two, δ_{DZ} . We use δ' to denote the above estimand, which is a function of the probability of treatment assignment at each level that depends on the experimental design. Whenever $\delta' \neq \delta_Z$, an unbiased estimator of δ' is a biased estimator of δ_Z .

We next show that, if researchers ignore the individual-level treatments regardless of the fact that q_0 and q_1 are not both zero in a joint multi-level design, the expected difference between the two potential outcomes corresponding to the two cluster-level treatments, $D = 1$ and $D = 0$, is likely to be an inappropriate estimand for the cluster-level treatment effect, δ_D .

$E[Y_{ij}^{(D_j=1)} - Y_{ij}^{(D_j=0)}] = \delta_D + (q_1 - q_0)\delta_Z + q_1\delta_{DZ}$. (14) The above expected difference would be the correct estimand for the cluster-level treatment effect δ_D only under one of the following two conditions: (a) $q_0 = q_1 = q$, and $\delta_{DZ} = 0$, or (b) $\delta_Z = \delta_{DZ} = 0$. In other words, the individual-level treatment assignment and treatment effect must not depend on the cluster-level treatment, or the individual-level treatment must have no effect whatsoever. When neither of these two conditions is satisfied, a sample mean difference between the two treatment groups, $D = 0$ and $D = 1$, would be a biased estimator of the cluster-level treatment effect δ_D .

To conclude, mistaking a joint multi-level design for either a multi-site design or a cluster design would very likely lead to misinterpretations of the estimated sample statistics, and would distort our understanding of the individual- and organization-level treatment effects.

Propensity Score-Based Approaches to the Causal Inference for Multi-Level, Observational Data

When we have observational rather than experimental multi-level data, a major challenge is to sort out the potential sources of bias. To clarify the sets of covariates that may need to be adjusted for in multi-level observational data, we use X to denote observed individual-level covariates, U_X for unobserved individual-level covariates, W for observed cluster-level covariates, and U_W for unobserved cluster-level covariates. Here W and U_W might include the first two moments of X and U_X at the cluster level. We define the propensity score(s) to be estimated for each type of multi-level observational design, and explore the propensity score-based approaches to the causal inference for each design.

In a *multi-site observational study* that resembles a multi-site randomized trial, different treatments are assigned to individuals within each site. Selection bias may come from both the individual and the cluster levels. In other words, X , U_X , W , and U_W may predict Z . Consequently, Z is likely to depend on the potential outcomes. Nonetheless, if we control for cluster effects via a fixed effects model, or if we center Z around its cluster mean in a linear model setting, Z becomes independent of the potential outcomes Y given X and U_X . Then we only need to worry about possible selection bias associated with the individual-level covariates. The conditional probability for individual i in cluster j to be assigned to the experimental group can be expressed as follows.

$$P_{ij} = \Pr(Z_{ij} = 1 \mid X_{ij}, U_{Xij}, j). \tag{15}$$

Given j , Z is independent of W_j and U_{Wj} . Moreover, when the strong ignorability assumption holds, U_X and U_W are no longer associated with Z given X and W . We estimate the propensity score P_{ij} as either a function of X_{ij} and W_j or a function of $(X_{ij} - \bar{X}_{\cdot j})$.

In a *cluster observational study* that resembles a cluster randomized trial, treatments are assigned at the cluster level. Potential bias is associated with only the cluster-level covariates W and U_W . Therefore, the cluster-level treatment assignment D is independent of X , U_X , and potential outcomes Y given W and U_W . The conditional probability of assigning cluster j to the experimental condition is as follows.

$$Q_j = \Pr(D_j = 1 \mid W_j, U_{Wj}). \tag{16}$$

The same probability applies to all the individuals within cluster j . Under the strong ignorability assumption, U_W is no longer associated with D given W . Hence, the propensity score Q_j can be estimated as a function of W_j only.

Next we examine a typical *joint multi-level observational design*. The observational data involve

a set of binary treatments at each of two different levels. The cluster-level treatment assignment D is likely to be associated with cluster-level covariates W and U_W ; while the individual-level treatment assignment Z may be associated with X and U_X that are centered around their respective cluster means. Following the notation that we used for a joint multi-level randomized design, Q_j denotes the conditional probability of assigning cluster j to the cluster-level experimental condition, and q_{0ij} and q_{1ij} denote the conditional probabilities of assigning individual i in cluster j to the individual-level experiment when cluster j is under the control condition or the experimental condition, respectively.

$$\begin{aligned} Q_j &= \Pr(D_j = 1 \mid W_j, U_{Wj}); \\ q_{0ij} &= \Pr(Z_{ij} = 1 \mid D_j = 0, X_{ij}, U_{Xij}, j); \\ q_{1ij} &= \Pr(Z_{ij} = 1 \mid D_j = 1, X_{ij}, U_{Xij}, j). \end{aligned} \tag{17}$$

Correspondingly, we obtain estimates of the following three propensity scores under the strong ignorability assumption.

$$\begin{aligned} \hat{Q}_j &= \Pr(D_j = 1 \mid W_j); \\ \hat{q}_{0ij} &= \Pr(Z_{ij} = 1 \mid D_j = 0, X_{ij}, j); \\ \hat{q}_{1ij} &= \Pr(Z_{ij} = 1 \mid D_j = 1, X_{ij}, j). \end{aligned} \tag{18}$$

One way to proceed with causal inference is to first match the clusters under the same treatment condition on the basis of \hat{Q}_j . Then within each pair of matched clusters, the individual experimental units and control units are to be matched on the individual-level propensity scores \hat{q}_{0ij} or \hat{q}_{1ij} , dependent upon the cluster-level treatment.

Propensity score stratification can be applied as an alternative adjustment strategy. We first divide all the clusters into, say, five strata on the basis of \hat{Q}_j . Within each stratum of clusters, we pool together all the individual units from the control clusters, and stratify them on the basis of \hat{q}_{0ij} . Similarly, we pool together all the individual units from the experimental clusters, and stratify them on the basis of \hat{q}_{1ij} . Assuming additivity, we can adopt a model-based approach to the treatment effect estimation. For individual i in cluster j , the combined model can be written as follows.

$$\begin{aligned} Y_{ij} &= (1 - D_j) [(\gamma_{00} + \gamma_{01} Z_{ij} + \sum_{s=1}^4 \gamma_{s+1,0} L_{sij}) \\ &+ D_j [(\gamma_{10} + \gamma_{11} Z_{ij} + \sum_{s=1}^4 \gamma_{s+1,1} L_{sij}) + \sum_{t=1}^4 \gamma_{0,t+1} M_{tj} \\ &+ \sum_{t=1}^4 \sum_{s=1}^4 \gamma_{st} L_{sij} M_{tj} + u_{0j} + u_{1j} Z_{ij} + e_{ij}; \\ u_{0j} &\sim N(0, \tau_{00}); (1 - D_j) u_{1j} + D_j u_{1j} \sim N(0, \tau_{11}); \\ \text{Cov}(u_{0j}, u_{1j}) &= \tau_{01}; e_{ij} \sim N(0, \sigma^2). \end{aligned} \tag{19}$$

Here L_{sij} , $s = 1, \dots, 4$, are dummy indicators for four of the five strata that stratify individuals in the control clusters or the experimental clusters. Note that the L series is a combination of two separate sets of propensity score strata, one for the control clusters, and the other for the experimental clusters. The dummy indicators for four of the five cluster-level strata are M_{ij} , $t = 1, \dots, 4$. We use γ_{0l} to denote the model-based estimator of δ_Z ; $\gamma_{10} - \gamma_{00}$ as an estimator of δ_D ; and $\gamma_{11} - \gamma_{01}$ as an estimator of δ_{DZ} .

We can also explore the application of the IPTW method to a joint multi-level observational design. For example, we can construct a cluster-level weight that is a function of the cluster-level treatment that a cluster actually received, and an individual-level weight that is a function of the individual-level treatment received. By using these weights to create pseudo populations at both levels, we expect to remove selection bias in the observational data.

Application Study: Effects of Kindergarten Retention and Kindergarten Retention Policy

Most educational treatments are carried out in school settings. The effectiveness of an individual-level intervention often depends on the organizational context within which the intervention takes place. In this application study we use the Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K) data to investigate the causal effects of kindergarten retention on children's cognitive growth in literacy and mathematics. The causal questions to be addressed include not only the effects of retention versus promotion as individual-level treatments, but also the effects of retention versus non-retention as school-level policies on children's learning.

The population of interest is comprised of students who are at risk of repeating the kindergarten year. There are three options for these at-risk students: being retained in a retention school, being promoted in a retention school, and being promoted in a non-retention school. We conceptualized the causal problems under two different scenarios. In the first and relatively simplified scenario, as almost all the previous retention researchers did, we tentatively ignored the school-level retention policy, and viewed retention as an individual-level treatment and retention assignment as resembling a hypothetical multi-site randomized design. At-risk students in each school are assigned to be either retained in kindergarten or be promoted to the first-grade. In the second scenario, we considered simultaneously the effect of the school-level retention policy and that of the individual-level retention treatment. Here we resorted to a hypothetical joint multi-level randomized design. Schools are assigned to two different policy conditions: retention being allowed at

the kindergarten level versus no kindergarten retention. In schools where retention is legitimate, students who are at risk of repeating are assigned to retention or promotion treatments.

The data contain repeated observations of a nationally representative sample of students, their families, teachers, and schools over the kindergarten and first-grade years. Our analytic sample included 1,117 promoted students in 141 non-retention schools, and 471 kindergarten retainees and 10,255 promoted students in 1,080 retention schools.

In our first scenario, the estimand for the retention effect is the expected difference between the potential outcome associated with retention and that associated with promotion. A single propensity score is needed to represent the treatment assignment mechanism. We identified a total of 207 individual-, classroom-, or school-level pretreatment variables that show statistically significant bivariate association with the individual-level treatment assignment. After replacing the missing values with their predicted values, we specified the logit of one's propensity of receiving the retention treatment as a function of 38 predictors, the quadratic terms for 17 of these 38 variables, and one interaction term. By excluding 774 promoted students who do not have counterparts in the retained group, we empirically identified a sample of children at risk of kindergarten repetition. The reduced sample included 12,746 students nested in 4,321 classrooms and 1,714 schools.

In order to adjust for pretreatment selection bias, we adopted the causal inference method that combines propensity score stratification with covariance adjustment. This was accomplished through estimating a 3-level model with individual students at level 1, treatment-year classrooms at level 2, and treatment-year schools at level 3. We also estimated the between-school variation in the retention effect. The population average retention effect on reading achievement was estimated to be -9.67 with a standard error of 0.64. We found statistically significant variation in the retention effect on reading between schools (estimated variance = 19.74, $\chi^2 = 269.66$, $df = 232$, $p < .05$). With a normal distribution of the retention effect assumed for the population of schools, the retention effect on reading would range from -18.38 to -0.96 among 95% of the schools. In addition, we found a correlation of -0.34 between school mean reading achievement and school-specific retention effect. Retention seems to do more harm to the kindergarten retainees in schools in which the average reading achievement is higher.

The retention effect on math was estimated to be -6.84 with a standard error of 0.50. This retention effect also varies significantly across

schools (estimated variance = 14.45, $\chi^2 = 287.02$, $df = 232$, $p < .01$), ranging from -14.29 to 0.61 among 95% of the schools. Contrary to the case in reading, the correlation between school mean math achievement and school-specific retention effect on math is positive with a magnitude of 0.49, suggesting a more detrimental retention effect in schools with a lower average math achievement.

In the second scenario, the estimand for the individual-level treatment effect is the expected difference between the potential outcome associated with retention in a retention school and that associated with promotion in a retention school. The estimand for the school-level policy effect is the expected difference between the potential outcome associated with promotion in a retention school and that associated with promotion in a non-retention school. Two different propensity scores need to be estimated, one for the school-level retention policy adoption, and the other for the student-level retention treatment assignment for students enrolled in retention schools. We identified 238 pretreatment covariates that are associated with the retention policy. Our final school-level propensity model included 29 covariates and the quadratic terms for 4 of these 29 variables. For students enrolled in the retention schools, after examining 207 pretreatment covariates that are associated with the individual-level retention treatment, we specified a propensity model that included 47 pretreatment covariates, quadratic terms for 10 of these 47 covariates, and one interaction term. To be consistent, we excluded from our subsequent analyses the same 774 promoted students that were excluded from the first-scenario analyses.

We stratified on the propensity scores and adjusted for covariance at each level. Based on our exploratory analysis, we included only the main effects of the propensity score strata. The estimated average difference between the potential outcomes associated with kindergarten retention and promotion, respectively, for a student in a retention school is -9.80 points with a standard error of 0.65. This estimated retention effect on reading amounts to about 70% of one standard deviation of the reading outcome. Meanwhile, being promoted in a retention school may bring some slight advantage compared with being promoted in a non-retention school. A student who is promoted in a retention school is expected to achieve 1.76 points more in reading, about 13% of one standard deviation of the reading outcome. This estimated difference is significantly different from zero, and is to be attributed to the school-level retention policy. We found no statistically significant variation in the retention effect on reading across the retention schools.

As to the retention effect on math, a child who is retained is expected to gain 6.25 points less than otherwise it would be if the same child is promoted in a retention school, almost 70% of one standard deviation of the math outcome. Interestingly, the retention effect on math varies across the retention schools (estimated variance = 13.41; $\chi^2 = 271.6$, $df = 232$, $p < .05$), and co-varies with school mean outcome. The correlation between school-specific retention effect and school mean math outcome is as high as 0.63. Kindergarten retention shows a more severe negative effect in schools in which the average math achievement is lower. Similar to the case in reading, the school-level retention policy shows some slightly positive effect, 1.25 points, which is about 14% of one standard deviation of the math outcome.

On the basis of the above substantive findings, we conclude that, in general, kindergarten retention impedes a child's cognitive growth in both reading and math. Interestingly, the results that we have obtained from the second scenario seem to suggest that the best option for an at-risk child is to be promoted in a retention school in which some of the similarly at risk peers are retained. This poses a difficult dilemma for policy makers and educators.

We intend to use this application study to illustrate different ways of conceptualizing a design for multi-level observational data and to compare their consequent results. According to our previous discussion, when we mistook a joint multi-level design for a multi-site design, as we did in the first scenario, the estimand would be inappropriately defined. In this application, let us denote that $\Pr(D = 1) = Q$, $\Pr(Z = 1 \mid D = 1) = q$, and $\Pr(Z = 1 \mid D = 0) = 0$. Equation 13 can be simplified as follows.

$$E[Y_{ij}^{(Z_{ij}=1)} - Y_{ij}^{(Z_{ij}=0)}] = \delta = \delta_Z + \delta_D \frac{1-Q}{1-qQ}. \quad (20)$$

We obtained an estimate of δ in the first scenario, and estimated both δ_Z and δ_D in the second scenario. The difference between δ and δ_Z depends on the magnitude of δ_D and q . In our case, the probability of kindergarten repetition in a retention school is no more than 0.05. Because q is relatively small, so is the retention policy effect δ_D , the deviance of our estimate of δ from that of δ_Z seems to be trivial. Nonetheless, we note that, by adopting the joint multi-level design that more closely matches the observed reality, we were able to reduce the residual variance in school-specific retention effect as well as in school mean achievement. For example, the residual variance in school retention effect on reading is 19.71 in the result of the first scenario and is reduced to 17.20 in that of the second scenario. The correlation between school mean achievement and

school-specific retention effect changes accordingly for both reading and math. Hence, the importance of choosing an appropriate multi-level design in analyzing multi-level observational data should not be overlooked.

Notes

¹Rubin (1978) argued that different versions of a treatment should be considered as different treatments. This is an important advice to follow when there are qualitative or quantifiable distinctions between treatment versions under the same treatment label. However, in observational studies of social practices, the number of versions of a treatment can be as many as the number of agents who deliver it. If such variation cannot be readily identified and perhaps scaled along certain dimensions, the convention has been to leave them in the error term.

²We do not attempt to exhaust all types of multi-level designs in this paper. For example, we will not discuss experiments in which individual units are assigned at random to treatments that are conducted in group settings. Experiments that assign units to treatments regardless of their clustering are not included in our discussion either. Nonetheless, the general principles that we illustrate in this paper can be adapted to these different designs.

References

- Cook, K. S. (1977). Exchange and power in networks of inter-organizational relations. *Sociological Quarterly*, 18, 62-82.
- Emerson, R. (1962). Power-dependence relations. *American Sociological Review*, 27, 31-40.
- Lindley, D. V. (1972). *Bayesian Statistics, A Review*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-41.
- Lipsky, M. (1980). *Street-level bureaucracy : dilemmas of the individual in public services*. New York: Russell Sage Foundation.
- Manski, C. F., & Garfinkel, I. (1992). *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. (Vol. 27). New York: Oxford University Press.
- Pfeffer, J., & Salancik, G. R. (1978). *The external control of organizations: A resource dependence perspective*. New York: Harper & Row.
- Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13(2), 85-116.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological methods*, 5(2), 199-213.
- Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (pp. 69-117). New York: Springer Verlag.
- Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 95-134). New York: Springer.
- Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1024-1032.
- Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of Royal Statistical Society, B*, 53(3), 597-610.
- Rosenbaum, P. R. (2002). *Observational Studies*. (2nd ed.). New York: Springer.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516-524.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34-58.
- Rubin, D. B., & Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, 79(4), 797-809.
- Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52, 249-264.
- Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustment for prognostic covariates. *Journal of the American Statistical Association*, 95(450), 573-585.