Estimation of Means and Covariance Components in Multi-site Randomized Trials

(DRAFT)

Stephen W. Raudenbush

Department of Sociology, Harris School of Public Policy,
and the Committee on Education
University of Chicago

Daniel Schwartz
Department of Statistics
University of Chicago

July 13, 2016

Abstract

In a multi-site randomized trial, units are assigned at random to treatments within sites such as schools, hospitals, neighborhoods, or cities. Such studies are now common in education, social welfare, and medicine. Although most analyses to date have focused on estimating the average impact of random assignment, a richer summary includes the variance of impacts across sites and the correlation between the site-specific impact and the site-specific control group mean. The precise meaning of these quantities will depend on the population of interest. In a two-level setting, for example, one may generalize population of persons or a population of sites. This article introduces consistent estimation of means and covariance components of site-specific intention-to-treat effects in multi-site trials using a procedure that maximizes a weighted, multi-level, normal-theory log-likelihood function. Weights that are functions of population sizes and sample sizes at each level are tailored to the target population of interest. We combine these with inverse probability of treatment weights within a hierarchical model (IPTW-HM) to eliminate biases that arise from unequal treatment allocation across sites. The data thus weighted typically emulate a family of balanced experimental designs, in which case our estimators constitute a class of nn-iterative method-of-moments estimators the properties of which do not depend on normality assumptions. While unbiased, these estimators may be inefficient. We compare this approach to two currently available analytic strategies for multi-site trials, each of which uses precision weighting: a site fixed effects (FE) model that uses ordinary least squares to estimate the average treatment effect; and a model with site fixed intercepts and random coefficients (FIRC) that estimates the mean and variance of treatment effects. We prove that FIRC produces a smaller bias and smaller standard errors than does FE for estimating the average treatment effect defined over a population of sites. We provide approximations for the FIRC bias and variance for estimating the average impact and the variance of the impacts across sites, and we show how to use these seek the best method for analyzing the data. We illustrate these methods by re-analyzing data from two major multi-site trials: US National Welfare to Work experiment and the US National Head Start Impact Study.

1. Introduction

Since 2002, the US Institute for Education Sciences (IES) has funded over 175 large-scale randomized trials to study the impact of a wide range of innovations. The vast majority of these trials are "multi-site" trials (Spybrook, 2013); that is, within each of many sites, units are assigned at random to participate in a novel program. In education, schools may be randomized within districts, classrooms may be randomized within schools, and sometimes the children themselves are randomized within classrooms or schools. However, in all cases, the multi-site trial is a "fleet of replicated experiments," each conducted in a specific site characterized by specific local organizational conditions and clientele. This paper focuses on methods for studying the average impact, the variation in impact across sites, and the correlation between the site-specific control group mean and the site-specific impact of a new intervention.

In this paper we confine our interest to studies in which the sites may be regarded as representing a universe of sites that might have been selected for the study or that might contemplate adopting the intervention of interest. Hence, of the studies listed in Table 1, the National Head Start Impact Study is clearly relevant, as the sites in that study constituted a formal probability sample from the universe of Head Start sites.[1] We regard the Tennessee Class Size Reduction Experiment as also relevant. In that study, children and teachers were assigned at random to a large or small class within each of 79 schools conceived here as sites. Although the 79 schools were not a formal probability sample, the planners of this study clearly aimed to generalize the results of this study to a larger universe of schools in Tennessee, with the aim of shaping statewide policy. However, the Moving To Opportunity Study (Kling, Liebman, and Katz, 2007), while extraordinarily important, falls outside the class of cases we shall consider because the sites were 5 large US cities – too few to be regarded as a sample from a larger universe of cities and therefore best regarded as "fixed" rather than "random" sites. Nor will we explicitly consider multi-site quasi-experiments, such as Chicago's "Double-Dose Algebra" trial of a new math curriculum in each of 60 high schools (Allensworth and Nomi, 2009) because in that study students were selected for the novel intervention not at random but rather by scoring below the cut point on a math test, yielding a "Regression Discontinuity Design" (Cook and Campbell, 1979). However, our findings can readily be extended to a wide range of multi-site quasi-experiments; these can be regarded as approximations of a randomized trial for sub-sets of persons who are similar in background.

[1] HSIS did not perfectly implement the ideal paradigm as not all sites yielded diversity in treatment, and there was substantial non-compliance such that a non-negligible number of children offered a place in Head Start and a substantial number of children assigned to the control condition managed to find a place in Head Start. Nevertheless, HSIS is remarkable in its capacity to produce an experiment that generalizes to a well-defined target population. The current paper does not consider the problem of compliance.

Table 1: Some Recent MS Trials

Study	Levels	Assigned Units	Sites	Fixed or Random sites
National Head Start Eval.	2	Children	378 Program Sites	Random
Moving to Opportunity	2	Families	5 cities	Fixed
Boston Charter School Lotteries	2	Children	Lottery pools	Random
Tennessee STAR	3	Teachers	Schools	Random
4 R's	3	Classrooms	Schools	Random
Double-Dose Algebra	2	Children	Schools	Random

Optimal sample size and treatment allocation will depend on the inferential target. In the Head Start Study, the aim, at least implicitly, was to generalize to the population of all Head Start Centers. This implies that if the Centers were a simple random sample, and if we knew the true impact of Head Start in every site sampled, we would weight these site-specific impacts equally in estimating the population average. The variance of the treatment effect would be of interest, as would the correlation between the control group mean and the treatment effect. This correlation would tell us whether sites that serve children whose prognosis without Head Start is poor are sites that tend to produce large (or small) treatment effects. Barring differences in cost or variance across sites, the optimal design would have a constant sample size per site with a constant fraction of children allocated to treatment. In practice, sample sizes varied quite dramatically from site to site in the Head Start study, apparently as a function of how many families applied to Head Start, and the fraction of persons assigned to Head Start varied accordingly, given the limited budget in each site.

Had the experimenters wished to generalize to a population of children, a sensible plan would have made the sample size per site proportional to the size of the eligible child population in that site, holding constant across sites the fraction assigned to treatment (again assuming constant site-specific costs and variances).

For simplicity, we restrict our interest to studies in which there are two treatment arms in each site, though extension to multi-valued treatments is straightforward. Optimal design for multi-site trials depends on the cost of sampling at each level, the magnitude of variation within treatments at each site and across sites, and the inferential goal (Raudenbush and Liu, 2000). The methods proposed here are readily adaptable to unequal costs and variances across sites, but for simplicity, the exposition here will focus on the case where costs and variances do not vary across sites. Covariates may be

introduced to increase precision or to model heterogeneity in treatment impact, but we avoid consideration of covariates to focus on key ideas.

We'll show that, under weak assumptions, combining design weights with inverse probability of treatment weights within a hierarchical linear model (IPTW-HM) estimated by maximizing a weighted log-likelihood function provides consistent estimates of means and covariance components in multisite trials. However, this approach may produce less precise estimates than do approaches that weight site-specific estimates by their precisions. To examine the resulting bias-variance tradeoff, we compare ITPW-HM to two currently available estimation strategies.

The first and most widely used approach uses ordinary least squares (OLS) regression with site fixed effects (FE) to estimate the overall average impact. The approach restricts interest to the average treatment effect. The appeal of this method is that it uses only the within-site information to estimate the average effect. It therefore eliminates possible confounding between site-specific treatment allocation and/or sample size with site-specific intercepts. However, the FE model assumes a homogeneous treatment effect. If site-specific impacts are heterogeneous and are correlated with site-specific treatment allocation or sample size, FE will produce a bias estimate of the overall treatment effect and an inflated variance.

Bloom, Raudenbush, Weiss, and Porter (2015) introduced an alternative approach with site-specific fixed effects and random coefficients (FIRC). This method shares the key virtue of FE: It uses only within-site information to estimate the average treatment effect. However, FIRC provides an estimate of the variance of the impacts across sites. We'll prove that FIRC is non-strictly less biased than is FE for the mean treatment effect when the aim is to generalize to a population of sites and the sites in the study are regarded as a simple random sample of sites. In contrast FE is non-strictly less biased for the mean treatment effect than is FIRC when the the aim is to generalize to a population of persons and the persons in the sample are regarded as a self-weighting sample of persons in the population.

Section 2 defines potential outcomes and person-specific causal effects, leading naturally to a simple theoretical model for the two-level multi-site trial. We define estimands that depend on the target population of interest and derive weights that identify these estimands. Section 3 introduces consistent estimation by maximizing a weighted multilevel log likelihood. Section 4 compares these estimators to FE and FIRC. Section 5 reanalyzes data from the National Welfare to Work experiment and National Head Start Study. We select these studies because they are substantively important and because they provide an interesting contrast in sampling plans. We quantify the variation in impact across sites and asks whether sites with high control group means produce large or small treatment effects. Using data from these studies, we show how to use our approximations to assess the asymptotic relative efficiency of FE, FIRC, and IPTW-HM for estimating the overall average treatment effect; and the asymptotic relative efficiency of FIRC and IPTW for estimating the variance of the treatment effects. We also study IPTW-HM's estimates of the covariance between the control group mean and the treatment effect.

Finally, we extend the IPTW approach to three-level trials, which arise commonly in educational experiments.

2. Theoretical Model and Estimands

2.1 Theoretical Model

For clarity of exposition, we refer to level-1 units here as persons and the level-2 units as sites. We invoke the Stable Unit Treatment Value Assumption (Rubin, 1986), under which each participant possesses one and only one potential outcome under each treatment condition. If person *i* within site *j* is assigned to treatment ($T_{ij} = 1$), we will observe the outcome $Y_{ij}(1)$; if that person is assigned to control ($T_{ij} = 0$), we will observe $Y_{ij}(0)$. The forgoing definitions can be summarized in a simple two-level hierarchical model for potential outcome $Y_{ij}(t)$, $t \in \{0,1\}$ of participant *i* in site *j*:

$$\begin{aligned} Y_{ij}(t) &= tY_{ij}(1) + (1-t)Y_{ij}(0) \\ &= Y_{ij}(0) + B_{ij}t \\ &= U_{0j} + \beta_j t + e_{ij}(t) \end{aligned} \quad . \tag{2.1}$$

where $B_{ij} = Y_{ij}(1) - Y_{ij}(0)$ is the person-specific causal effect of assignment to treatment, $U_{0j} = E(Y_{ij}(0) \mid U_{0j})$ is the average untreated outcome for the SUB-population of persons in site *j*; $\beta_j = E(B_{ij} \mid \beta_j)$ is the average treatment effect for this sub-population; and $e_{ij}(t) = Y_{ij}(0) - U_{0j} + t(B_{ij} - \beta_j)$ is a random error. We shall assume that, this random error has variance

$$\begin{aligned} Var(e_{ij}(t)) &= Var[Y_{ij}(0) - U_{0j}] + t[Var(B_{ij}) + 2Cov(Y_{ij}(0), B_{ij})] \\ &\equiv \sigma_{0j}^2 + t(\sigma_{Bj}^2 + 2\sigma_{0Bj}). \end{aligned} \tag{2.2}$$

Looking across a population of sites, the site-specific control-group mean and treatment effect randomly vary

$$\begin{aligned} U_{0j} &= \mu_0 + u_{0j} \\ \beta_j &= \mu_\beta + b_j \end{aligned} \tag{2.3}$$

where the random effects u_{0j} and b_j have zero means and variance-covariance matrix

$$Var\begin{pmatrix} u_{0j} \\ b_j \end{pmatrix} = \begin{bmatrix} \tau_{00} & \tau_{0b} \\ \tau_{0b} & \tau_{bb} \end{bmatrix}. \tag{2.4}$$

2.1 Estimands

In a population composed of J^* sites each composed of a sub-population of N_j^* persons, we define the average treatment effect over a population of persons as

$$E_{persons}(B) = \sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} B_{ij} / N^* = \sum_{j=1}^{J^*}(N_j^* / \overline{N}^*)\beta_j / J^* \equiv \mu_{\beta\ persons} \tag{2.5}$$

where N^* is the total number of persons in the sub-population and $\overline{N}^* = N^* / J^*$ is the average sub-population size across sites.

In contrast, suppose we are interested generalizing to a population of sites (e.g., day care centers, schools, or hospitals), the population-mean impact will be

$$E_{sites}(B) \equiv \sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} \overline{N}^* / N_j B_{ij} / N = \sum_{j=1}^{J^*}\beta_j / J^* \equiv \mu_{\beta\ sites}. \tag{2.6}$$

We can regard (2.5) and (2.6) as special cases of the more general form

$$E_L(B) \equiv \sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}^L B_{ij} / N^* = \sum_{j=1}^{J^*} w_{2ij}^L \beta_j / J^* \equiv \mu_{\beta_L} \tag{2.7}$$

where *L* specifies the level to which we shall generalize and we are weighting person-specific causal effects B_{ij} with weights w_{1ij}^L or, equivalently, we are weighting site-specific impacts β_j with weights are w_{2j}^L. When generalizing to a population of persons as in Equation (2.5), we have

$$w_{1ij}^{persons} = 1 \text{ and } w_{2ij}^{persons} = N_j^* / \overline{N}^*. \tag{2.8}$$

In contrast, when generalizing to a population of sites (Equation 2.6), we have

$$w_{1ij}^{sites} = \overline{N}^* / N_j^*\ w_{2j}^{sites} = 1 \tag{2.9}$$

Clarifying the level of generalization and the appropriate weights at each level will be essential in clarifying estimands and estimators for two-level data and deriving the weights needed for estimation by IPTW-HM.

We shall focus on the first two moments of U_{0j}, β_j across sites, which have general form

$$\mu_{0_L} = \sum_{j=1}^{J^*} w_{2j}^L U_{0j} / J^* \qquad \mu_{\beta_L} = \sum_{j=1}^{J^*} w_{2j}^L \beta_j / J^*,$$

$$\tau_{00_L} = \sum_{j=1}^{J^*} w_{2j}^L (U_{0j} - \mu_0)^2 / J^* \qquad \tau_{BB_L} = \sum_{j=1}^{J^*} w_{2j}^L (\beta_j - \gamma)^2 / J^* \qquad (2.10)$$

$$\tau_{0B_L} = \sum_{j=1}^{J^*} w_{2j}^L (U_{0j} - \mu_0)(\beta_j - \gamma) / J^*,$$

where μ_{0_L} is the population-average control group mean and μ_{β_L} is the population-average impact; and τ_{00_L}, τ_{BBL}, and τ_{0B_L} are the population variance of the site-specific control group means, the population variance of the site-specific impacts, and the population covariance between the control group mean and impact. To reduce the notation, we shall often suppress the superscript and subscript "*L*" in the text that follows.

2.2 Identification

We cannot of course observe both $Y_{ij}(0)$ and $Y_{ij}(1)$. Instead, we observe $Y_{ij} = Y_{ij}(0) + B_{ij}T_{ij}$. Following Equations (2.1) and (2.3), we can write this observed outcome according to the mixed model

$$Y_{ij} = \mu_0 + \mu_\beta T_{ij} + u_{0j} + b_j T_{ij} + e_{ij}. \qquad (2.11)$$

where $e_{ij} = Y_{ij}(0) - U_{0j} + T_{ij}(B_{ij} - \beta_j)$. Because T_{ij} is randomly assigned, we confidently invoke the assumption of conditional independence of treatment assignment and potential outcomes within sites:

$$Y_{ij}(1), Y_{ij}(0) T_{ij} \perp T_{ij} \mid Site = j \qquad (2.12)$$

One of the key challenges in multi-site field trials is that we cannot assume independence of treatment assignment and potential outcomes marginally because the site-specific fraction of persons assigned to treatment, that is $\bar{T}_j = \sum_{i=1}^{n_j} T_{ij} / n_j$, will tend to vary from site to site, often non-ignorably. For example, many multi-site trials recruit participants to apply for a lottery and use the lottery to assign participants at random to a novel treatment. The number of people who apply and therefore the probability of winning the lottery may depend on endogenous factors such as the perceived popularity of the program in each site, availability of alternative programs in a site, and the intensity of participants' interest in treatment generally. In other studies, the number of available slots may depend on the resources available at each site or other factors not under control of the experimenter.

Identification of the mean impact. It's useful to define the "prima facie" causal estimands (Holland, 1986) for the two definitions of the target population. This motivates the derivation of the weights that identify the model for each target population.

$$\begin{aligned}\mu_{\beta pf} &= E(Y \mid T=1) - E(Y \mid T=0) \\ &= \mu_\beta + E(u_0 \mid T=1) - E(u_0 \mid T=0) + E(b \mid T=1) + E(e \mid T=1) - E(e \mid T=0) \\ &= \mu_\beta + E(u_0 \mid T=1) - E(u_0 \mid T=0) + E(b \mid T=1).\end{aligned} \tag{2.13}$$

Conditionally independent treatment assignment ensures that $E(e \mid T=t) = (e) = 0.$ The form of the bias in $\mu_{\beta pf}$ will depend on the target population through the definition of the weights:

$$\begin{aligned}Bias(\mu_{\beta pf}) &= E(u_0 \mid T=1) - E(u_0 \mid T=0) + E(b \mid T=1) \\ &= \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij} T_{ij} u_{0j}}{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij} T_{ij}} - \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij} (1-T_{ij}) u_{0j}}{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{ij} (1-T_{ij})} + \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij} T_{ij} b_j}{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij} T_{ij}} \\ &= J^{*-1} \sum_{j=1}^{J} w_{2j} \left(\frac{\overline{T}_j}{\overline{T}} - \frac{1-\overline{T}_j}{1-\overline{T}} \right) u_{0j} + J^{*-1} \sum_{j=1}^{J} w_{2j} \frac{\overline{T}_j}{\overline{T}} b_j \\ &= \frac{Cov(\overline{T}_j, u_{0j})}{\overline{T}(1-\overline{T})} + \frac{Cov(\overline{T}_j, b_j)}{\overline{T}}\end{aligned} \tag{2.14}$$

where we define $Cov(\overline{T}_j, v_j) \equiv \sum_{j=1}^{J^*} w_{2j} \overline{T}_j v_j / J^*$ for a site-level zero-man random variable v_j. Equation 13 reveals that we can identify the average causal effect by applying a level-1 weight that is the product of w_{1ij} tailored to target the desired population, and the IPTW, yielding

$$\omega_{1ij} = w_{1ij} \left[T_{ij} \frac{\overline{T}}{\overline{T}_j} + (1-T_{ij}) \frac{1-\overline{T}}{1-\overline{T}_j} \right]. \tag{2.15}$$

Identification of the level-2 covariance structure. Unbiased estimation of the site means and the overall mean insures identification of the level-2 covariance components; we simply make substitutions into Equation 10. In practice, modest sample sizes of persons per sites or of sites will pose challenges for estimation, as we'll see in the next section.

Identification of the level-1 variances. Identification of the level-1 variance structure is interesting in itself and essential for estimation of the level-2 covariance structure (see the next section). It's convenient to re-write the level-1 model as

$$
\begin{aligned}
Y_{ij}(t) &= tY_{ij}(1) + (1-t)Y_{ij}(0) \\
&= t(U_{1j} + e_{ij}(1)) + (1-t)(U_{0j} + e_{ij}(0)),
\end{aligned}
\tag{2.16}
$$

equivalent to Equation (2.3) with $U_{1j} = U_{0j} + \beta_j$, $e_{ij}(0) = Y_{ij}(0) - U_{0j}$, $e_{ij}(1) = Y_{ij}(1) - U_{1j}$, and $Var[e_{ij}(1)] \equiv \sigma_{1j}^2 = \sigma_{0j}^2 + \sigma_{Bj}^2 + 2\sigma_{0Bj}$. Next, we define pooled level-1 variances as

$$
\sigma_t^2 = \frac{\sum_{j=1}^{J^*} w_{2j}\sigma_{tj}^2}{\sum_{j=1}^{J^*} w_{2j}}, \quad t = 0,1. \tag{2.17}
$$

A prima facie estimand σ_{1pf}^2 has form

$$
\begin{aligned}
\sigma_{1pf}^2 = E[e_i^2 \mid T=1] &= \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}T_{ij}e_{ij}^2(1)}{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}T_{ij}} = \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}(T_{ij} - \overline{T}_j + \overline{T}_j)e_{ij}^2(1)}{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}(T_{ij} - \overline{T}_j + \overline{T}_j)} \\
\backslash \\
&= \frac{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}(T_{ij} - \overline{T}_j)e_{ij}^2(1) + \sum_{j=1}^{J^*} w_{2j}\overline{T}_j\sigma_{1j}^2}{\sum_{j=1}^{J^*}\sum_{i=1}^{N_j^*} w_{1ij}(T_{ij} - \overline{T}_j) + \sum_{j=1}^{J^*} w_{2j}\overline{T}_j}.
\end{aligned}
\tag{2.18}
$$

Because the potential outcomes are independent of *T* within sites, (2.18) tends to

$$
\sigma_{1pf}^2 = E[e_i^2 \mid T=1] = \frac{\sum_{j=1}^{J^*} w_{2j}\overline{T}_j\sigma_{1j}^2}{\sum_{j=1}^{J^*} w_{2j}\overline{T}_j} = \sigma_1^2 + Cov\left(\frac{\overline{T}_j}{\overline{T}}, \sigma_{1j}^2\right) \tag{2.19}
$$

as J^* increases without bound. We can readily see that using (2.15) as a weight eliminates the covariance term in (18) and hence identifies σ_1^2. Application of (2.17) similarly identifies σ_0^2. An important result is that heteroscedasticity across sites is no bar to estimation of the two level-1 variances so long as we specify separate overall level-1 variances σ_1^2 for treatment participants and σ_0^2 for controls. This fact becomes important in estimating the level-2 variance components as we shall see in the next section.

3. Consistent Estimation Using Inverse Probability of Treatment Weighting and Sample Design Weights within a Hierarchical Model

Pfefferman et al. (1997) proposed sample design weights for a two-level hierarchical linear model with random intercepts. We extend that approach to include random coefficients and, later, to three levels of variation. For the two-level case, we write down a weighted "complete-data" log likelihood and apply an EM algorithm to obtain weighted maximum likelihood estimates. When these weighted log-likelihood functions simulate the log-likelihood that would be obtained from a family of balanced designs, the estimators are generally equivalent to non-iterative method-of-moments estimators. Hence, such estimators will not rely on normality assumptions even though the "working model" (Meng, 2015) is a normal theory model. This broadens the range of outcomes that can be studied. However, such estimators may be quite inefficient in particular applications, an issue we consider later.

Robins, Hernan, and Brumback (2000) developed IPTW as a strategy for adjusting for confounding in non-experimental studies. Hong and Raudenbush (2008) embedded IPTW within a hierarchical HLM in their non-experimental assessment of math instruction. Hong (2010) developed theory for a non-parametric approach to IPTW embedded within a multilevel model; Hong and Hong (2009) and Hong, Corter, Hong, and Pelletier (2012) applied this approach in non-experimental settings. Application to multi-site randomized trial is particularly straightforward because each site's sample propensity score $\overline{T}_j$ is known, omitting one of the key problems in many applications of IPTW, namely the need to estimate the propensity score using data on observed covariates. Extensions to multiple levels is straightforward. A unique focus of the current paper is the definition of estimands at various levels and the integration of sample design weights and IPTW to identify means and covariance components defined for various target populations.

3.1 Defining Weights for Two Level Multi-site Trials for a Population of Persons

Let's first consider a two-level study in which persons are assigned at random within each of many sites. We'll define weights to accomplish two purposes. The first is to ensure that our estimators are tailored to the target population of interest. The second is to eliminate confounding between the sample propensity score $\overline{T}_j$ and site-level random effects.

Targeting the population of persons. Suppose that we could compute the true impact for each person in the sub-population of a given site. As described in the introduction (Equations 1-5), we'd weight each by 1.0 if we wanted to generalize to the population of persons; and that would be consistent with weighting the site-average effect by N_j / N. Combining this insight with IPTW gives the level-1 and level-2 weights when the target population is persons:

$$L = persons,\ \ w_{1ij} = \frac{N_j^*}{\overline{N}} * \frac{\overline{n}}{n_j} \qquad w_{2j} = \frac{N_j^*}{\overline{N}} \tag{3.1}$$

where $\overline{n} = \sum_{j=1}^{J} n_j$ and $\overline{N} = \sum_{j=1}^{J} N_j^*$. If persons were sampled with probability proportional to size, so that $n_j / \sum_{j=1}^{J} n_j = N_j^* / \sum_{j=1}^{J^*} N_j$, the weights would simplify, with $w_{1ij} = 1$.

In many studies, we do not have a probability sample of sites or persons. However, in this paper, we'll treat each as equally representative of an undefined "super-population" of sites that might adopt an intervention and each site's set of n_j sampled persons are regarded as representing a sub-population of persons in that site. In this case, the weights defined in (3.1) are appropriate. In the remainder of this article, we shall assume the existence of such a sub-population.We'll also generally regard the sample sizes as negligibly small fractions of the population sizes; thus we'll treat n_j / N_j^* and J / J^* as negligible. To eliminate bias and target the desired population we apply the level-1 weight (2.15) with $w_{1ij} = 1$.

3.2 Defining Weights for Two Level Multi-site Trials for a Targeting a Population of Sites

If we wish to generalize to population of sites (each site counted equally), we'd have weighted the site-average effects by 1.0. Combining this insight with IPTW gives the level-1 and level-2 weights when the target population is persons:

$$\omega_{1ij} = \frac{\overline{n}}{n_j}\left(T_{ij}\frac{\overline{T}}{\overline{T}_j} + (1 - T_{ij})\frac{1-\overline{T}}{1-\overline{T}_j} \right) \qquad w_{2j} = 1. \tag{3.2}$$

3.3 Maximizing a Two-Level Weighted Log-Likelihood

To obtain consistent estimators, we maximize a weighted two-level normal theory log likelihood. An appealing way to do this is to first write down the "complete data" weighted log-likelihood, that is, the weighted version of likelihood that we would maximize if the level-2 random effects were observed (Demptster, Rubin, and Tsutakawa, 1981). Let's translate our model into a more general representation of a two-level normal theory model,

$$Y_{ij} = X_{ij}^T(\alpha + a_j) + e_{ij}, \quad a_j \sim N(0, \tau),\ \ e_{ij} \sim N(0, \sigma_{ij}^2) \tag{3.3}$$

where $X_{ij}^T = \begin{pmatrix} 1 & T_{ij} \end{pmatrix}$, $\alpha = \begin{pmatrix} \mu_0 & \mu_\beta \end{pmatrix}^T$, $a_j = \begin{pmatrix} u_{0j} & b_j \end{pmatrix}^T$, τ is a symmetric, positive definite 2 by 2 matrix having unique elements $\tau_{00}, \tau_{0b,} \tau_{bb}$ and e_{ij} is an independently sampled level-1 random effect having variance

$$\sigma_{ij}^2 = T_{ij}\sigma_{1j}^2 + (1 - T_{ij})\sigma_{0j}^2. \tag{3.4}$$

We use the EM algorithm (Dempster, Laird, and Rubin, 1977) which simulates maximum likelihood on the "complete data" (Y, a) even though the random effects, a, are missing. The weighted complete-data log-likelihood is

$$\begin{aligned} \sum_{j=1}^{J}\sum_{i=1}^{n_j} l_{wij} = &-\tfrac{1}{2}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n_j} \omega_{1ij}[(Y_{ij} - X_{ij}^T\alpha - X_{ij}^T a_{rj})^2 / \sigma_{ij}^2 + \log(2\pi\sigma_{ij}^2)]\right\} \\ &-\tfrac{1}{2}\left[\sum_{j=1}^{J} w_{2j} a_j^T \tau^{-1} a_j + J\log(2\pi\,|\,\tau\,|)\right] \end{aligned} \tag{3.5}$$

where $\sum_{j=1}^{J} w_{2j} = J$, $\sum_{j=1}^{J}\sum_{i=1}^{n_j} \omega_{1ij} = \sum_{i=1}^{n_j} n_j \equiv N$.

EM operates by first defining the "complete-data maximum likelihood estimates" (CDMLE) and then estimating, on each iteration, the weighted "complete-data sufficient statistics" on which those CDMLE depend by their conditional expectations given the observed data *Y* and estimates of the parameters from the previous iteration. These estimated CDMLE are then substituted into the formulas for the CDMLE to obtain a new estimate of the parameter vector. In the case of the models using weights defined by (3.2) , the algorithm terminates after one iteration for any reasonable starting values because we are simulating a class of balanced designs in which the estimators have closed form.

3.4 Resulting Two-Level Estimators

Table 2 describes the estimating equations we obtain for generalizing to a population of persons and for generalizing to a population of sites. These estimators are unbiased and consistent for our estimands. See Appendix A for the derivation of these results.

Table 2: Weighting schemes and resulting estimating equations.

Target	Population of Persons[a]	Population of Sites
Level-1 weight	$T_{ij}\frac{\overline{T}}{\overline{T}_j}+(1-T_{ij})\frac{1-\overline{T}}{1-\overline{T}_j}$	$\frac{\overline{n}}{n_j}\left(T_{ij}\frac{\overline{T}}{\overline{T}_j}+(1-T_{ij})\frac{1-\overline{T}}{1-\overline{T}_j}\right)$
Level-2 weight	$n_j/\overline{n}$	1
Mean impact	$\hat{\mu}_\beta=\sum_{j=1}^{J}\frac{n_j}{\overline{n}}\hat{\beta}_{1j}/J$	$\hat{\mu}_\beta=\sum_{j=1}^{J}\hat{\beta}_{1j}/J$
Control group mean	$\hat{\mu}_0=\sum_{j=1}^{J}\frac{n_j}{\overline{n}}\hat{U}_{0j}/\sum_{j=1}^{J}n_j$	$\hat{\mu}_0=\sum_{j=1}^{J}\hat{U}_{0j}/J$
Variation in impact[b]	$\hat{\tau}_{bb}=J^{-1}\sum_{j=1}^{J}\frac{n_j}{\overline{n}}\left[(\hat{\beta}_{1j}-\hat{\mu}_\beta)^2-V_j\right]$	$\hat{\tau}_{bb}=J^{-1}\sum_{j=1}^{J}\left[(\hat{\beta}_{1j}-\hat{\mu}_\beta)^2-V_j\right]$
Correlation between impact and control mean[c]	$\hat{\tau}_{0b}=J^{-1}\sum_{j=1}^{J}\frac{n_j}{\overline{n}}\left[(\hat{\beta}_{1j}-\hat{\mu}_\beta)(\hat{\beta}_{0j}-\hat{\mu}_0)-C_j\right]$	$\hat{\tau}_{0b}=J^{-1}\sum_{j=1}^{J}\left[(\hat{\beta}_{1j}-\hat{\mu}_\beta)(\hat{\beta}_{0j}-\hat{\mu}_0)-C_j\right]$

[a]We assume here that that the sample size per site is proportional to the sub-population size in that site.

[b]The sampling variance $V_j=E(\hat{\beta}_j-\beta_j)^2=n_j^{-1}[\sigma_1^2/\overline{T}_j+\sigma_0^2/(1-\overline{T}_j)]$.

[c]The sampling covariance $C_j=E(\overline{Y}_{cj}-U_{0j})(\hat{\beta}_j-\beta_j)=-n_j^{-1}\sigma_0^2/(1-\overline{T}_j)$.

We see from Table 2 that all estimators are defineable in closed form. These are equivalent to method-of-moments estimators. This will not be the case when we include covariates or when, in the case of targeting a population of persons, site-sample sizes are not proportional to site population sizes.

3.5 Consistent Estimation for Three Level Multi-site Trials

The majority of randomized trials funded by the US Institute of Education Sciences have randomized clusters within sites (Spybrook, 2013). For example in some studies, students are nested within classrooms with schools and the classrooms are assigned at random to treatments. The schools might further be nested within districts. And indeed, the schools themselves might be randomly assigned to treatments.

We can readily generalize the strategy of embedding IPTW within a hierarchical linear model to any number of levels. We'll illustrate how this works in the case of three level trials. For concreteness, let's assume that children $i=1,\ldots,n_{jk}$ are nested within classrooms $j=1,\ldots,J_k$ within schools $k=1,\ldots,K$ and that classrooms are assigned at random to treatments. For simplicity, we'll assume further that each of the n_{jk} children sampled in classroom *j*,*k* equally represents a sub-population of children who might have

been assigned to that classroom; that each classroom has one teacher; that each of the J_k teachers in school *k* equally represents a sub population of teachers who might have been assigned to that school; and that each school equally represents a sub population of schools that might have been sampled for the experiment.

Possible target populations. We might wish to generalize results to a population of children, a population of teachers, or a population of schools. If we wish to generalize to a population of schools, we might conceive each school as a sub-population of students or a sub-population of teachers. So we must select one of these four target populations. Which target we select will determine how the weights would be constructed and applied to obtain estimators (see Table 3).

Table 3: Target Populations and Corresponding Weights for Three-Level Models

Target population	Students	Teachers	Schools (each school is a sub-population of students)	Schools (each school is a sub-population of students)
Level-1 weight	1	$\frac{\bar{n}}{n_{jk}}$	$\bar{J}\bar{n}/(J_k\bar{n}_k)$	$\bar{J}\bar{n}/(J_k n_{jk})$
Level-2 weight	$\frac{n_{jk}}{\bar{n}} * \left(\frac{T_{jk}\bar{T}}{\bar{T}_k} + \frac{(1-T_{jk})(1-\bar{T})}{1-\bar{T}_k}\right)$	$\frac{T_{jk}\bar{T}}{\bar{T}_k} + \frac{(1-T_{jk})(1-\bar{T})}{1-\bar{T}_k}$	$\frac{\bar{J}n_{jk}}{J_k\bar{n}_k} * \left(\frac{T_{jk}\bar{T}}{\bar{T}_k} + \frac{(1-T_{jk})(1-\bar{T})}{1-\bar{T}_k}\right)$	$\frac{\bar{J}}{J_k} * \left(\frac{T_{jk}\bar{T}}{\bar{T}_k} + \frac{(1-T_{jk})(1-\bar{T})}{1-\bar{T}_k}\right)$
Level-3 weight	$J_k\bar{n}_k/(\bar{J}\bar{n})$	J_k/J	1	1

Note:

$$\bar{n}_k = \sum_{j=1}^{J_k} n_{jk}/J_k \qquad \bar{n} = \sum_{k=1}^{K}\sum_{j=1}^{J_k} n_{jk} \qquad \bar{J} = \sum_{k=1}^{K} J_k/K$$

$$\bar{T}_k = \sum_{j=1}^{J_k} T_{jk}/J_k \qquad \bar{T} = \sum_{k=1}^{K}\sum_{j=1}^{J_k} T_{jk} \qquad T_{jk} \in \{0,1\}$$

We apply these weights to the model

$$\begin{aligned} Y_{ijk} &= X_{jk}^T(\alpha + a_k) + r_{jk} + e_{ijk}, \\ a_k &\sim N(0,\tau), \quad r_{jk} \sim N(0,\omega_r^2) \quad e_{ijk} \sim N(0,\sigma^2) \end{aligned} \tag{3.6}$$

where $X_{jk}^T = \begin{pmatrix}1 & T_{jk}\end{pmatrix}$, $\alpha = \begin{pmatrix}\mu_0 & \mu_\beta\end{pmatrix}^T$; $a_k = \begin{pmatrix}u_{0k} & b_k\end{pmatrix}^T$ is a vector of school-level random effects, r_{jk} is a scalar teacher random effect, and , τ is a symmetric, positive definite 2 by 2 matrix having unique elements $\tau_{00}, \tau_{0b,}\tau_{bb}$ and e_{ijk} is the level-1 random effect.

We again apply the EM algorithm (Dempster, Laird, and Rubin, 1977). The "complete" data now include Y, a, r with random effects a, r constituting the missing data. The weighted complete-data log-likelihood is

$$\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{i=1}^{n_{jk}} l_{wijk} = -\tfrac{1}{2}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{1ijk}[(Y_{ijk} - X_{tj}^{T}(\alpha + a_k) - r_{jk}]^2 / \sigma^2 + N\log(2\pi\sigma^2)\right]$$
$$-\tfrac{1}{2}\left[\sum_{j=1}^{J} w_{2jk} r_{jk}^2 / \omega^2 + J_k \log(2\pi\omega^2)\right] \qquad (3.7)$$
$$-\tfrac{1}{2}\left[\sum_{j=1}^{J} w_{3k} a_k^T \tau^{-1} a_k + K\log(2\pi\,|\,\tau\,|)\right]$$

$$(where \sum_{k=1}^{K} w_{3k} = K, \quad \sum_{k=1}^{K}\sum_{j=1}^{J_k} w_{2jk} = J \qquad \sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_j} w_{1ijk} = N).$$

4. Alternative Estimators that Use Precision Weighting

The beauty of IPTW-HM is that, under weak assumptions, it provides consistent estimation of each of means and covariance components in our theoretical model. However, these estimates may produce large sampling variances as compared to competing estimators. In this section, we consider two competitors. The first is the widely used site fixed effects (FE) estimator. The second is a hybrid estimator that uses fixed intercepts and random coefficients (FIRC). We compare these approaches in the context of a two-level multi-site trial.

FE uses a form of precision-weighting to estimate the average treatment effect while FIRC uses a different form of precision weighting to estimate the average treatment effect and the variance. Bias will arise if the precision weights are associated with random, site-specific random terms. Specifically, suppose that each site generates an independent, unbiased estimate $\hat{\theta}_j$ of a parameter μ_θ with known precision ϕ_j and that we estimate μ_θ by means of the precision weighted average

$$\hat{\mu}_\theta = \sum_{j=1}^{J} W_j \hat{\theta}_j / J \qquad (4.1)$$

where $W_j = \phi_j / \bar{\phi}$, with $\bar{\phi} = \sum_{j=1}^{J} \phi_j / J$. Assuming that W_j is effectively random (that is, not fixed by design), such an estimator will have bias

$$E(\hat{\mu}_\theta - \mu_\theta) = Cov(W_j, \hat{\theta}_j) = \frac{\sum_{j=1}^{J^*} w_{2j}^L (W_j - 1)\hat{\theta}_j}{\sum_{j=1}^{J^*} w_{2j}^L}. \qquad (4.2)$$

We can therefore compare alternative precision-weighted averages by comparing these covariance terms. We can derive mean squared errors of competing estimators by combining this information with information about sampling variances.

4.1 Site Fixed Effects (*FE*)

Economists have tended to prefer a site fixed effects (*FE*) model (Greene, 2003; Angrist and Pischke, 2008) to random coefficient models. The standard *FE* model assumes constant treatment effects, so interest is confined to the population average treatment effect, effectively assuming $\tau_{00} = \tau_{bb} = \tau_{0b} = 0$. The model is

$$Y_{ij} = \mu_\beta T_{ij} + \alpha_j + e_{ij} \tag{4.3}$$

where α_j is a site-specific fixed effect and e_{ij} is a within-site random error. This leads to the pooled, within-site ordinary least squares estimator

$$\hat{\mu}_{\beta FE} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(T_{ij} - \overline{T}_j)(Y_{ij} - \overline{Y}_j)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(T_{ij} - \overline{T}_j)^2} = \frac{\sum_{j=1}^{J} W_j^{FE}\hat{\beta}_j}{J}. \tag{4.4}$$

where $\hat{\beta}_j = \overline{Y}_{1j} - \overline{Y}_{0j}$, and $W_j^{FE} = P_j / \overline{P}$ with $P_j = 1/Var(\hat{\beta}_j \mid \hat{\beta}_j)$ and $\overline{P} = \sum_{j=1}^{J} P_j^{FE} / J$. If one assumes homogeneous level-1 variance (that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$), $P_j = n_j \overline{T}_j (1 - \overline{T}_j) / \sigma^2$. However, this assumption can readily be relaxed.

4.2 Site Fixed Effects with Random Coefficients (*FIRC)*

Bloom, Raudenbush, Weiss, and Porter (2009) expanded the conventional fixed effects model by incorporating a site-specific random treatment effect. This enabled them to estimate the average treatment effect and the variance of the treatment effects but not the correlation between the treatment effect and the control group mean. The resulting "fixed intercepts random coefficient model" (FIRC) can be written

$$Y_{ij} = (\mu_\beta + b_j)T_{ij} + \alpha_j + e_{ij}, \tag{4.5}$$

where α_j is again a fixed constant while $b_j \sim N(0, \tau_{bb})$. Estimation requires an iterative method. If we apply the method of iterative generalized least squares (Goldstein, 1986) for the mean and variance of the treatment effects, we obtain at iteration *m*+1 the estimating equations

$$\hat{\mu}_{\beta FIRC}^{(m)} = \sum_{j=1}^{J} \Delta_j^{(m-1)-1}\hat{\beta}_j \Big/ \sum_{j=1}^{J} \Delta_j^{(m-1)-1} \tag{4.6}$$

and

$$\tau_{bb}^{(m)} = \sum_{j=1}^{J} \Delta_j^{(m-1)-2}[(\hat{\beta}_j - \hat{\mu}_{\beta_{FIRC}}^{(m-1)})^2 - V_j]/\sum_{j=1}^{J} \Delta_j^{(m-1)-2}, \quad (4.8)$$

where $\Delta_j^{(m-1)} = \tau_{bb}^{(m-1)} + V_j$ where $V_j = 1/P_j$. (For $\tau_{bb}^{(m)} \le 0$, set $\tau_{bb}^{(m)} = 0$, in which case $\hat{\mu}_{\beta_{FIRC}}^{m+1} = \hat{\mu}_{\beta_{FE}}$.) When $\tau_{bb}^{(\infty)} > 0$, a particularly useful form of (4.6) is

$$\hat{\mu}_{\beta_{FIRC}} = \sum_{j=1}^{J} W_j^{FIRC} \hat{\beta}_j / J \quad (4.9)$$

where $W_j^{FIRC} = \lambda_j / \bar{\lambda}$, where λ_j is the estimated "reliability of $\hat{\beta}_j$ as a measure of β_j (Raudenbush and Bryk, 2002, Chapter 3) and $\bar{\lambda} = \sum_{j=1}^{J} \lambda_j / J$. Specifically

$$\lambda_j = \tau_{bb}^{(\infty)} / (\tau_{bb}^{(\infty)} + P_j^{-1}) \quad (4.10).$$

4.3 Comparing the Bias of $\hat{\mu}_{\beta_{FE}}$ and $\hat{\mu}_{\beta_{FIRC}}$

Theorem 1: If the aim of the study is to generalize to a population of sites, each weighted equally, $Bias(\hat{\mu}_{\beta_{FIRC}}) \le Bias(\hat{\mu}_{\beta_{FE}})$.

Proof: <<The Proof is Forthcoming >>

5. Comparing the Weighted Estimator to Its Competitors

Embedding sample design weights and IPTW within the framework of a hierarchical linear model enables us to estimate all of the parameters of that model consistently. Indeed, this seems to be the only available method that does so without making strong assumptions.

In the pursuit of consistency, however, a bias-variance tradeoff arises. The IPTW-HM approach removes bias, but the bias we remove by weighting may be very small while the added variance we thus generated may be comparatively large.

In this section, we compare the asymptotic efficiency (that is, efficiency as the number of sites *J* increases without bound) of the weighting approach to that of alternative approaches when interest is confined to the average treatment effect and the variance of the treatment effects. To keep the presentation as simple as possible we'll restrict our attention to two-level multi-site trials, and we'll focus on the case in which the target population is a population of sites. It is in this case where the estimators are most likely to diverge.

5.1 Sampling Precision is Independent of Impact

Let's first consider the efficiency of competing estimators of the average treatment effect when site-specific sampling precisions, which are based on propensity scores and sampling precisions, are ignorable. In this case, all of estimators considered here are unbiased for the parameters they estimate. We'll see that the IPTW approach can be quite inefficient in these cases. Table 4 displays the sampling variances of the three approaches when sampling precisions are independent of site-specific impacts.

First, if the sample sizes and/or propensity scores vary but $\tau_{bb} = 0$, FE and FIRC are asymptotically identical and more efficient than IPTW, a fact that becomes clear when we inspect the ratio of variances given $\tau_{bb} = 0$:

$$\frac{Var(\hat{\mu}_{\beta_{IPTW},\tau_{bb}=0})}{Var(\hat{\mu}_{\beta_{FE},\tau_{bb}=0},)} = \frac{Var(\hat{\mu}_{\beta_{IPTW},\tau_{bb}=0})}{Var(\hat{\mu}_{\beta_{FIRC},\tau_{bb}=0},)} = \frac{E\left(P^{-1}\right)}{[E(P)]^{-1}} \equiv \frac{E(P)}{\mu_{hp}} \tag{5.1}$$

where $E(P)$ is the arithmetic mean of the sampling precisions P_j and μ_{hp} is the harmonic mean of these precisions. This ratio can be considerably greater than unity when the precisions vary a lot, as we'll see in the illustrative example below.

Table 4: Sampling variances of three estimators when the aim is to generalize to a sites

	Estimator	**Sampling Variance**
OLS Site Fixed Effects (FE)[a]	$\hat{\mu}_{\beta_{FE}} = \frac{\sum_{j=1}^{J} P_j \hat{\beta}_j}{\sum_{j=1}^{J} P_j}$	$\frac{1}{J\overline{P}} + \frac{\tau_{bb}}{J}\left(1 + \frac{Var(P)}{\overline{P}^2}\right)$
Fixed Intercepts Random Coefficients (FIRC)[b]	$\hat{\mu}_{\beta_{FIRC}} = \frac{\sum_{j=1}^{J} (\tau_{bb} + P_j^{-1})^{-1} \hat{\beta}_j}{\sum_{j=1}^{J} (\tau_{bb} + P_j^{-1})^{-1}}$	$\frac{1}{J * E(\tau_{bb} + P^{-1})^{-1}}$
Random Intercepts and Coefficients with weighting (IPTW)	$\hat{\mu}_{\beta_{FE}} = \frac{\sum_{j=1}^{J} \hat{\beta}_j}{J}$	$\frac{E(\tau_{bb} + P^{-1})}{J}$

and site-specific sampling precisions and site-specific treatment effects are independent.

a,b **The** expressions for the sampling variances for FE are asymptotic in the number of sites *J*.

Second, when sample sizes and propensity scores are ignorable and $\tau_{bb} > 0$, FIRC dominates IPTW; that is, the asymptotic relative efficiency of FIRC compared to IPTW is, given by

$$\frac{Var(\hat{\mu}_{\beta_{IPTW}})}{Var(\hat{\mu}_{\beta_{FIRC}})} = \frac{E\sum_{j=1}^{J}\left(\tau_{bb} + P_j^{-1}\right)}{E\left\{\left[\sum_{j=1}^{J}\left(\tau_{bb} + P_j^{-1}\right)^{-1}\right]^{-1}\right\}} \equiv \frac{E(\overline{\Delta})}{E(\tilde{\Delta})} \geq 1 \qquad (5.2)$$

where $\overline{\Delta}$ is the arithmetic mean of $\Delta_j = Var(\hat{\beta}_j) = \tau_{bb} + P_j$ and $\tilde{\Delta}_j$ is the corresponding harmonic mean. Hence, we conclude that when sample sizes and propensity scores are ignorable, FIRC dominates HM-IPTW as well as FE.

Finally, we know from theory that when sample sizes and propensity scores are ignorable and $\tau_{bb} > 0$, FIRC dominates FE. Specifically, in the model $\hat{\beta}_j = \mu_\beta + E_j, \quad E_j \sim N(\gamma, \Delta_j)$, $\hat{\mu}_{\beta_{FIRC}}$ achieves the Cramer-Rao lower variance bound in large-*J* samples. Without normality, $\hat{\mu}_{\beta_{FIRC}}$ is best linear unbiased in large samples. Approximations to the asymptotic relative efficiency are complicated, but we can easily evaluate the formulas in Table 1 in applications, as we shall illustrate.
Recall also that FIRC also dominates FE (see last section).

5.2 Sampling Precision and Impact are Related

What happens when sample sizes and propensity scores are *not* ignorable? In this case, $\hat{\mu}_{\beta_{IPTW}}$ remains unbiased, while $\hat{\mu}_{\beta_{FE}}$ and $\hat{\mu}_{\beta_{FIRC}}$ are biased and inconsistent. Hence, in infinite samples, $\hat{\mu}_{\beta_{IPTW}}$ is more efficient than either of the other two methods. However, in moderate to large samples, the biases may be small compared to the variance reduction associated with $\hat{\mu}_{\beta_{FIRC}}$ and even $\hat{\mu}_{\beta_{FE}}$. Therefore, to select a method of estimation, we'd like to know how large these biases are. We know from Section 4 that $Bias_{FIRC} \leq Bias_{FE}$, but we'd like to assess the plausible size of these biases in application.

We can achieve these goals approximately by representing the two biases as a Taylor series. Expanding the Taylor series in the natural log precision $\eta_j = \ln(P_j)$ around the geometric mean $\tilde{P}$ of $P_j, j = 1,...,J$ works well. To obtain a first-order approximation, we'll treat η_j as normally distributed, that is $\eta_j \sim N(\bar{\eta}, \sigma_\eta^2)$ where $\bar{\eta} = \ln(\tilde{P})$. Appendix B (forthcoming) will provide a more general approximation. Under this set up, we'll see that

$$Bias(\hat{\mu}_{\beta\,RIRC}) \cong E\,\frac{\sum_{j=1}^{J} b_j[\hat{\lambda} + \hat{\lambda}(1-\hat{\lambda})]\eta_j}{J\hat{\lambda}} = \sigma_{\eta b}(1-\hat{\lambda}) \qquad (5.3)$$

where $\sigma_{\eta b} = Cov(\eta, b)$, $\hat{\lambda} = \tau_{bb}/(\tau_{bb} + \hat{P}^{-1}) < 1$, and we require that $\sigma_\eta^2 = Var(\eta_j) < 1$ for the approximation to hold, a constraint that we find very realistic in application (see our illustrative examples). For comparison, we see that

$$Bias_{FE} = Cov(P/\bar{P}, b) \cong \sigma_{\eta b}. \qquad (5.4)$$

The approximations are intuitively meaningful. If $(1-\bar{\lambda})$ is zero, which will occur when impacts are highly heterogeneous, that is τ_{bb} is large relative to the sampling variance $V_j = P_j^{-1}$, FIRC weights, proportional to $\Delta_j^{-1} = (\tau_{bb} + \sigma^2 P_j^{-1})^{-1}$, will be nearly homogenous and $\hat{\mu}_{\beta_{FIRC}} \approx \hat{\mu}_{\beta_{IPTW}}$; hence the FIRC bias will be null. In contrast, $(1-\bar{\lambda})$ will approach 1.0 when τ_{bb} is small relative to the sampling variance $V_j = \sigma^2 P_j^{-1}$, so FIRC weights will approximate those of FE and the FIRC bias will approach the FE bias as well. We'll see how to use these ideas in the illustrative example.

Table 5 summarizes these results. It also provides expressions for the asymptotic variance of the estimators of mean effect size. We derived the FE and FIRC variances under the assumption that *b* and $\eta = \ln(P)$ are bivariate normal. Appendix B provides our derivations and provides more complex expressions that are needed if the bivariate normal assumption is abandoned.

Table 5: Asymptotic Mean and Variance for the Average Treatment Effect

	Estimator	**Asymptotic Bias**[a]	**Asymptotic Variance**[b]
OLS Site Fixed Effects (FE)	$\hat{\mu}_{\beta_{FE}} = \frac{\sum_{j=1}^{J} P_j \hat{\beta}_j}{\sum_{j=1}^{J} P_j}$	$Cov(\eta, b)(1+\nu)$	$\frac{1}{J\overline{P}} + \frac{\tau_{bb}^{*}}{J}\left(1 + \frac{Var(P)}{\overline{P}^2}\right)$
Fixed Intercepts Random Coefficients (FIRC)	$\hat{\mu}_{\beta_{FIRC}} = \frac{\sum_{j=1}^{J} (\tau_{bb} + P_j^{-1})^{-1} \hat{\beta}_j}{\sum_{j=1}^{J} (\tau_{bb} + P_j^{-1})^{-1}}$	$(1-\overline{\lambda})Cov(\eta, b)(1+\varepsilon)$ $(0 < \overline{\lambda} < 1)$	$\frac{1}{J * E(\tau_{bb}^{**} + P^{-1})^{-1}}$
Random Intercepts and Coefficients with weighting (IPTW)	$\hat{\mu}_{\beta_{FE}} = \frac{\sum_{j=1}^{J} \hat{\beta}_j}{J}$	0	$\frac{E(\tau_{bb} + P^{-1})}{J}$

[a] $\overline{\lambda} = \tau_{bb}(1-\rho_{\eta,b}^2)/[\tau_{bb}(1-\rho_{\eta,b}) + \widetilde{P}^{-1}]^{-1}$ and $\widetilde{P}$ is the geometric mean of P.

$\overline{P} = E(P)$

ν, ε are errors of approximation (Appendix B)

[b] $\tau_{bb}^{*} = \tau_{bb}(1-\rho_{\eta,b}^2) + \sigma_{\eta B}^2$ where $\rho_{l\eta,b} = Corr(\eta, b)$

$\tau_{bb}^{**} = \tau_{bb}(1-\rho_{\eta,b}^2) + (1-\hat{\lambda})\sigma_{\eta B}^2$.

Table 5 shows that, to the first order, $Var(\hat{\beta}_{IPTW-HM}) \geq Var[\hat{\beta}_{FIRC} | \ln(P)]$ and $Var(\hat{\beta}_{FE}) \geq Var(\hat{\beta}_{FIRC})$.

5.2 Asymptotic Relative Efficiency for the Effect Variance

Now let's first take a look at the efficiency with which we can estimate the cross-site variance in treatment impacts. Now FE is out of the picture, because FE assumes homogeneous impacts. The story now is quite simple, because the FIRC estimate of the variance has negligible bias. To see this, let's approximate the expected variance estimate using the same method just described for the mean. We'll see then that the large-sample expectation of the FIRC variance estimator is

$$\frac{1}{E(\tau_{bb}^{**} + P^{-1})^{-2}}. \qquad (5.5)$$

The asymptotic efficiency of FIRC relative to *IPTW* is, to the first order

$$\frac{Var(\hat{\tau}_{bb,IPTW})}{Var(\hat{\tau}_{bb,FIRC})} = \frac{2\sum_{j=1}^{J}\Delta_j^2 / J}{2/\sum_{j=1}^{J}\Delta_j^{**-2}} \cong \frac{E(\Delta^2)}{\mu_{h\Delta^{**2}}} \geq 1 \qquad (5.7)$$

where $E(\Delta^2)$ is the arithmetic mean of the squared variance Δ and $\mu_{h\Delta^2}$ is the harmonic mean of Δ_j^{**-2}. The former will be considerably larger than the latter in many applications.

Table 5: Asymptotic Mean and Variance for the Variance of the Treatment Effect

	Estimator	**Asymptotic Bias**[a]	**Asymptotic Variance**[b]
Fixed Intercepts Random Coefficients (FIRC)	$\hat{\tau}_{bbFIRC} = \left[\sum_{j=1}^{J}\Delta_j^{-2}(\hat{\beta}_j - \hat{\mu}_\beta)^2 - \hat{V}_j\right] / \sum_{j=1}^{J}\Delta_j^{-2}$	≈ 0	$2/\sum_{j=1}^{J}\Delta_j^{**-2}$
Random Intercepts and Coefficients with weighting (IPTW)	$\hat{\tau}_{bbIPTW} = \left[\sum_{j=1}^{J}(\hat{\beta}_j - \hat{\mu}_\beta)^2 - \hat{V}_j\right] / J$	0	$2\sum_{j=1}^{J}\Delta_j^2 / J$

6. Illustrative Example

The Welfare to Work study is a useful example for the purposes of this paper. Sample sizes and propensity scores vary dramatically across the 59 sites of this study. Heterogeneity of effect is comparatively small across sites, but the large within-site sample sizes compensate so that the average reliability ($\bar{\lambda} = .412$) of site-specific impact estimates is respectable. This is important for our purposes because FIRC and FE would provide similar results if reliability were small. If the reliability were near 1.0, FIRC and IPTW would behave very similarly, enabling us to learn little.

6.1 Data

Our data set consists of pooled data from three multi-site trials conducted by MDRC over more than a decade: The Greater Avenues for Independence (GAIN) project conducted in 22 local welfare offices from six California counties (Riccio and Friedlander, 1992); Project Independence conducted in 10 local welfare offices from nine Florida counties (Kemple and Haimson, 1994); and the National Evaluation of Welfare-to-Work Strategies conducted in 27 local welfare offices from seven states (Hamilton, 2002). The goal was to enable persons on welfare to obtain jobs and earn more money.

Our aim here is to illustrate the logic of estimation theory for multi-site trials rather than to draw conclusions from the data. See Bloom et al. (2015) for substantive results, which we thank those authors for making available to us here. Table 6a provides the relevant data for our evaluation. Note the large and highly variable site sample sizes ($\bar{n} = 718,\ sd = 618$), and the highly variable propensity scores $\bar{T} = .49,\ sd = .13$; as a result, the sampling precision weights are also highly variable (note the standard deviation of 147.78). Despite this enormous variation, the variance of the log-transformed precisions of .79 is well under 1.0, the limit allowed by our approximations. Indeed, the approximations described above worked well as adding more terms in the Taylor series than those described above had a negligible effect. Table 6b shows that the impact variance estimate of 54902 =771^2 (in dollars) provided by Bloom et al. is substantively meaningful relative to the average impact of 770 dollars earned per year and was highly statistically significant despite its small size in comparison to the within-site variance.

Table 6a: Welfare to Work Design Features

	Minimum	Maximum	Mean	Standard deviation
Sample size n_j	137	2972	718	618
Propensity score $\bar{T}_j$	.49	.86	.66	.13
$P_j = n_j\bar{T}_j(1-\bar{T}_j)$	23.96	645.63	152.42	147.78
$\omega_j = \ln(P_j)$	3.17	6.47	4.68	0.79
Reliability, λ_j	.13	.80	.41	.18

Table 6b: Parameter estimates from past research

Parameter	Estimate
Average impact, μ_β	770
Within-site variance, σ^2	9715^2
Cross-site impact variance τ_{bb}	771^2

6.2 Results

Bias. We can approximate the ratio of bias of FIRC relative to FE for the average impact even without knowing the two biases using second-order approximations:

$$\frac{Bias_{FIRC}}{Bias_{FE}} \cong \frac{(1-\bar{\lambda})\{1+[\frac{1}{2}-3\bar{\lambda}(1-\bar{\lambda})]\hat{\sigma}_\omega^4\}}{1+(1-\bar{\lambda})(1-2\bar{\lambda})\hat{\sigma}_\omega^2(1+\frac{1}{2}\hat{\sigma}_\omega^2)} \qquad (34)$$
$$= 0.61$$

Variance. We estimate the efficiency of FIRC relative to IPTW as

$$\frac{Var(\hat{\mu}_{\beta_{IPTW}})}{Var(\hat{\mu}_{\beta_{FIRC}})} = \frac{\hat{E}\left(\tau_{bb} + \sigma^2 P^{-1}\right)}{\hat{E}\left(\tau_{bb} + \sigma^2 P^{-1}\right)^{-1}} = \frac{29261}{24410} = 1.20 \qquad (35)$$

and the efficiency of FIRC relative to FE as

$$\frac{Var(\hat{\mu}_{\beta_{FE}})}{Var(\hat{\mu}_{\beta_{FIRC}})} = \frac{\hat{E}\left(\frac{\sigma^2}{J\overline{P}} + \frac{\tau_{bb}}{J}\left(1 + \frac{Var(P)}{\overline{P}^2}\right)\right)}{\hat{E}\left(\tau_{bb} + \sigma^2 P^{-1}\right)^{-1}} = \frac{30029}{24410} = 1.23. \quad (35)$$

Interestingly, $Var(\hat{\mu}_{\beta_{IPTW}})$ and $Var(\hat{\mu}_{\beta_{FE}})$ are estimated to be nearly equivalent, suggesting that FE puts too much weight on the sampling precisions P_j. This is not too surprising given that $\overline{\lambda} \cong .41$while the FE weights are optimal when $\overline{\lambda} \cong 0$ and the IPTW approach of unit weighting is optimal when $\overline{\lambda} \cong 1$.

The relative efficiency of FIRC compared to FE for estimating the variance is approximately=1.51.

References

Angrist, J.D., and Pischke, J. (2008) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D., Imbens, G. and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association*, 91(434): 444-455.

Bloom, H. S. (1984) Accounting for No-Shows in Experimental Evaluation Designs, *Evaluation Review* 8(2): 225 – 46.

Bloom, H.S., Raudenbush, S.W., Weiss, M., and Porter, K.E (2015). *Using Multi-site Experiments to Study Cross-site Variation in Effects of Program Assignment*. MDRC.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1-38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*(374), 341-353.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education India, 2003.

Hamilton, G. (2002). *Moving people from welfare to work: Lessons from the National Evaluation of Welfare-to-Work Strategies*. Washington DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and U.S. Department of Education, Office of the Under Secretary and Office of Vocational and Adult Education.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46: 1251-1271.

Heckman, J.H., Pinto, R. and Savelyev (2013). Understanding the mechanisms through which an influuential early childhood program boosted adult outcomes. *American Economic Review 2013, 103(6): 2052–2086*

Hill, J. (2013). Multilevel models and causal inference in Scott, M., Simonoff, J., and Marx, B. (Eds). *The SAGE Handbook of Multilevel Modeling.*

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data Scott, M., Simonoff, J., and Marx, B. *Journal of Educational and Behavioral Statistics*, **35**(5), 499-531.

Hong, G., Corter, C. Hong, Y. & Pelletier, J. (2012). Differential effects of literacy instruction time and homogeneous grouping in kindergarten: Who will benefit? Who will suffer? *Educational Evaluation and Policy Analysis*, 34(1), 69-88.

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis*, 31(1), 54-81.

Hong, G., & Raudenbush, S.W. (2008). Causal inference for time-varying instructional treatments. *The Journal of Educational and Behavioral Statistics*, 33(3), 333-362.

Horvitz, D.l G., and Thompson, D.V. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47.260 (1952): 663-685.

Kemple, J. & Haimson, J. (1994). *Florida's Project Independence: Programimplementation, participation patterns and first-year impacts*. New York: MDRC.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83-119.

Mundlak, Y. (1978). On the pooling of time series and cross-sectional data. *Econometrica*, 46, 69-86.

Nomi, T., & Allensworth, E. (2009). "Double-Dose" Algebra as an Alternative Strategy to Remediation: Effects on Students' Academic Outcomes. *Journal of Research on Educational Effectiveness*, *2*(2), 111-148.

Pfefferman, C.J., Skinner, D.J., Holmes, H., Goldstein, H., Rashbash, J. (1998). Weighting for unequal selection probabilities in multilevel analysis. *Journal of the Royal Statistical Society, Series B,* 60,1, 23-40.

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Journal of Education, Finance and Policy.* Vol. 4, No. 4, pp 468 – 491.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological methods*, *5*(2), 199.

Riccio, J. & Friedlander, D. (1992). *GAIN: Program strategies, participation patterns and first-year impacts in six counties*. New York: MDRC.

Robins, J., Hernan, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560.

Rubin, D. B. (1986). Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, 81, 961–962.

Spybrook, Jessaca (2013) "Detecting Intervention Effects Across Context: An Examination of the Precision of Cluster Randomized Trials," *The Journal of Experimental Education*, DOI: 10.1080/00220973.2013.813364.

Spybrook, J. & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis.* Vol. 31, No. 3, pp 298 – 318.

US Department of Health and Human Services, & US Department of Health and Human Services. (2010). Head Start impact study: Final report. *US Department of Health and Human Services. Administration for Children and Families.(2004). Making a difference in the lives of infants and toddlers and their families: The impacts of Early Head Start*, *1*.

VanderWeele, T.J., Hong, G., Jones, S. and Brown, J. (2013). Mediation and spillover effects in group-randomized trials: a case study of the 4R's educational intervention. *Journal of the American Statistical Association*, 108:469-482.

Weiss, M.J., Bloom, H.S., and Brock, T. (2013). *A Conceptual Framework for Studying Sources of Variation in Program Effects. MDRC Working Papers in Research Methodology*. New York: MDRC.