An Applied Researcher's Guide to Estimating Effects From Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates

Online Appendix A: Technical Details for Estimands, Estimators, and Estimates Paper

July 13, 2020

1 Introduction

In this document we systematically discuss each of the various methods for estimating the ATE we consider for multi-site experiments. We in particular discuss how to calculate standard errors, along with possible pitfalls in doing so. We also unpack some of the intuition behind some of the estimators, such as how the weights in the weighted approaches give unbiased estimators. We also discuss connections between the estimators, such as how inverse weighting will give identical estimates to the design-based estimator and the fully interacted linear model.

When randomization of individuals occurs within multiple blocks within site (e.g., if there are multiple cohorts that were each randomized independently), there can be additional complications. For example, the definition of a site averaged effect is now implicitly an average over multiple blocks. For clarity we, in this section, assume that there is only one randomization block per site and refer to each group of units randomized to treatment and control as a site. See our separate discussion in Section 7.2 for further discussion of the complications of randomization blocks. We also assume that each site has multiple (at least two) treatment and control units, allowing for variance estimates for each treatment arm for each site. In the case where some sites have singleton units in the treatment or control arm, one has to take alternative steps; see **?** for further discussion and overview of the literature.

For notation, we follow the main text. In particular, we use the following notation throughout:

- β - The estimands, which are all average treatment effects (ATEs). They are subscripted such as $\beta_{SP-Persons}$ or $\beta_{FP-Sites}$.

- B_j and p_j - The ATE and the proportion treated in site j. We have J total sites, and index them with j.
- N_{1j} and N_{0j} - The total numbers of units receiving treatment and control in site j.
- T_{ij} - The treatment indicator for unit i in site j.
- N, N_1, N_0, and p - The Ns are the total number of units, total number of treated units, and total number of control units across all sites, and $p = N_1/N$ is the proportion of the entire sample treated.
- b_{rj}, n_{rj}, p_{rj} - The ATE, total number of units, and proportion of units treated for randomization block r in site j, when we are considering multiple randomization blocks nested within site.
- For superpopulation totals, we add a star such as N_j^* for total number of individuals in the full population of site j.

Before discussing the estimators themselves, we first further discuss finite sample versus super population inference. We then discuss the estimators in four major parts: design based, linear regression, linear regression with treatment by site interactions, and multilevel modeling. We finally discuss the further issues of estimation.

2 The relationship of finite-sample vs. superpopulation inference

In this section we discuss two fundamental observations. First, in terms of the estimates themselves, any estimator can be thought of as either a finite-sample or superpopulation estimate. Second, the true uncertainty of an estimate, when taking it as a superpopulation estimate, is the combination of the finite-sample uncertainty coupled with the variability of the entire experimental sample with respect to the superpopulation. This means it is only the uncertainty quantification, not the choice of point estimate, that needs to be explicitly tailored towards a finite-sample or superpopulation estimand.

To illustrate this, we first argue that the best estimate of a super population parameter is the corresponding finite sample parameter (which of course is not directly observed), assuming there is no extra population information under consideration. This immediately gives that result that the best obtainable estimator for the super population will be the best estimator we have for the finite population, and vice-versa. We then use this observation to explain why estimating the finite sample parameter will generally be more precise than the corresponding superpopulation parameter.

To begin, recall that our superpopulation framework assumes the sites are a simple random sample of the sites in the superpopulation. If the sample is a simple draw from the superpopulation, the sample of sites is representative and we have

$$\mathbb{E}\left[\beta_S\right] = \beta$$

i.e., the expected value of our finite-sample estimand will be the superpopulation estimand. As a consequence, for any finite-sample unbiased estimator we have, using the tower property of expectations,

$$\mathbb{E}\left[\hat{\beta}\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\beta}|S\right]\right] = \mathbb{E}\left[\beta_S\right] = \beta$$

where the "$\mathbb{E}\left[\cdot|S\right]$" indicates conditioning our expectation on a given (random) sample. We also have

$$\begin{aligned}\text{var}\left[\hat{\beta}\right] &= \mathbb{E}\left[\text{var}\left[\hat{\beta}|S\right]\right] + \text{var}\left[\mathbb{E}\left[\hat{\beta}|S\right]\right] \\ &= \mathbb{E}\left[SE^2\left[\hat{\beta}|S\right]\right] + \text{var}[\beta_S]\end{aligned}$$

where $SE^2\left[\hat{\beta}|S\right]$, the finite sample uncertainty, is the true finite-sample standard error of our estimator across the assigment of treatment, given a specific sample S. The above derivation shows several things.

First, it shows that the best unbiased estimator for β is β_S. This comes from the variance term: if we somehow had β_S as an estimator (i.e., $\hat{\beta} = \beta_S$) then the first term in the variance will drop out. The second term is independent of estimator, and thus we have the smallest variance.

This also shows that any unbiased estimator for the finite sample β_S is an unbiased estimator for the superpopulation β. Similarly, any unbiased estimator for the superpopulation estimand will be unbiased, across samples and randomization, for the finite.

The above also shows how the performance (in terms of the Standard Error or RMSE) of an estimator can differ depending on whether we view our estimate as estimating the finite sample estimand or superpopulation estimand. The variance decomposition, above, makes this more precise: the superpopulation uncertainty is a combination of the uncertainty of estimating the parameter in the finite sample and the uncertainty of the finite sample as representative of a larger population.

Valid estimates of superpopulation uncertainty should be close to $\text{var}\left[\hat{\beta}\right]$, which includes the $\text{var}[\beta_S]$ term. Standard errors for finite sample inference, on the other hand, will hover, across different samples drawn from the superpopulation, around the $\mathbb{E}\left[\text{var}\left[\hat{\beta}|S\right]\right]$ term, which will tend to be smaller. Furthermore, if there is no cross-site heterogeneity, i.e., $\text{var}[\beta_S] = 0$, then our finite population and superpopulation variances are the same: we only need to worry about uncertainty in generalizing if it is possible for a sample to have a different

ATE than the population. In this case, any finite sample is, by definition, representative of the superpopulation, in terms of average impact.

As a thought experiment to illustrate, consider the case of three sites. If the sites were suitably large, we may have nearly perfect knowledge of the average impact for each site and therefore for our sample as well. However, especially if the three site average impacts were quite different, we may have little knowledge of the average impact in the super population as we, in effect, have a sample of size 3. We would thus want our stated uncertainty to be quite different depending on which estimand we are targeting. This is why it is important to identify whether a given analysis is focused on the finite sample or infinite population: it is due to uncertainty estimation.

Importantly, there are actually several potential super populations we might appeal to. Borrowing from the categories in **?** and **?** we have the following:

1. Sites are fixed for the study, but individuals are randomly sampled within the sites from site-specific super populations.
2. Individuals with attached site membership are randomly sampled from a larger population, with site as a categorical covariate akin to race or gender.
3. Sites are randomly sampled from a super population of sites, but the individuals are fixed within site.
4. Sites are randomly sampled, and individuals within sites also randomly sampled within site.

The third is what we generally assume in our main text. The first two are actually more "finite population" as they do not allow for a broader concept of site variation: our sample gives a full representation of all the sites and we thus do not need to infer about a larger population of sites. The second of the above is what primarily motivates heteroskedastic robust standard errors. The super population targeting estimation strategies can differ in terms of which model they are assuming. Cluster robust standard errors assume the fourth, and multilevel models most closely adhere to the fifth.

3 Design Based Estimators

The design based estimators we consider are discussed extensively in **?**. They all work with the estimated impacts within each site of

$$\hat{B}_j = \hat{Y}_j(1) - \hat{Y}_j(0) \tag{1}$$

$$= \frac{1}{N_{1j}} \sum_{i=1}^{N_j} T_{ij} Y_{ij}(1) - \frac{1}{N_{0j}} \sum_{i=1}^{N_j} (1 - T_{ij}) Y_{ij}(0) \tag{2}$$

The site vs. individual estimators just reweight these site-level estimates differently.

The individual estimator is the simplest estimator where we take the weighted average of the estimated average impacts of the sites:

$$\hat{\beta}_{DB-persons} = \sum_{j=1}^{J} \frac{N_j}{N} \hat{B}_j. \tag{3}$$

This is the estimator classically used for blocked experiments or post-stratified experiments. See **?** and **?**.

To get our site average impact, simply average the individual site average impact estimates:

$$\hat{\beta}_{DB-site} = \sum_{j=1}^{J} \frac{1}{J} \hat{B}_j \tag{4}$$

The standard errors differ by finite population and superpopulation. We describe the two versions of standard error estimation next.

3.1 Design based finite population standard errors

The uncertainty estimates for finite sample design based estimators are all of the same structure, which is a weighted sum of site-level uncertainty estimates. Let w_j be a vector of weights across sites (e.g., either N_j/N for individual weighting or $1/J$ for site weighting). Then

$$\widehat{SE}\left[\hat{\beta}_{DB-w-finite}\right] = \left[\sum_{j=1}^{J} \frac{w_j^2}{W^2} \widehat{SE}^2\left[\hat{B}_j\right]\right]^{1/2} \text{ with } W = \sum_j w_j \tag{5}$$

with

$$\widehat{SE}^2\left[\hat{B}_j\right] = \frac{1}{N_{1j}} s_{1j}^2 + \frac{1}{N_{0j}} s_{0j}^2.$$

The s_{zj}^2 are the sample variances of the individuals assigned to treatment z in site j. Note the W is a normalizing constant; if the weights are N_j/N or $1/J$, then $W = 1$; this normalizing constant is why we could instead give the weights as simply $w_j = N_j$ for person-weighting or $w_j = 1$ for site-weighting. These uncertainty estimates are known to be conservative due to the correlation of potential outcomes problem. **?** proposes a slightly tighter bound, subtracting $(s_{1j} - s_{0j})^2/N_j$ from the expression.

When the estimand is the effect for the average site in a finite population, the $1/J$ weights allow for an algebraic simplification, giving

$$\widehat{SE}\left[\hat{\beta}_{DB-site-finite}\right] = \left[\frac{1}{J^2} \sum_{j=1}^{J} \widehat{SE}^2\left[\hat{B}_j\right]\right]^{1/2}$$

Importantly, any specified weighting of sites w_j can be used. In particular, if we have randomization blocks within site, these formula can be directly applied to the randomization blocks with weights tuned to target average site impacts. See Section 7.2, below.

3.2 Design based superpopulation standard errors

Finite-sample design based estimators traditionally focus on the assignment mechanism, but for superpopulation estimators we also have to account for the additional uncertainty of the sampling of the sites. This can depend on which specific sampling framework is used; see **?** for a discussion of the different superpopulation models possible under a blocking view.

In brief, for superpopulation design-based estimators, the treatment effect estimators themselves are the same as for the finite sample case. It is the uncertainty estimates that are now different. In the revised version of the RCT-Yes software we have the following:

$$\widehat{SE}^2\left[\hat{\beta}_{DB-w-super}\right] = \frac{1}{(J-1)J\bar{w}^2}\sum_{j=1}^{J} w_j^2\left(\hat{B}_j - \hat{\beta}\right)^2. \tag{6}$$

The above measures the dispersion of the site-average impacts across sites. These $\hat{B}_j$ will vary for two reasons: the within-site uncertainty due to the random assignment and the between-site uncertainty due to whether the sampled sites are representative of the superpopulation. Because these sources of error are independent, the uncertainty directly adds. To see the connection, take the simplest case of the site-average. Here we estimate $\hat{\beta}$ with the average of the $\hat{B}_j$, and the variance in this estimation is simply the estimated variance of the point estimates, divided by the sample size of the number of sites; this is the classic standard error for estimating a mean from a simple random sample:

$$\widehat{SE}^2\left[\hat{\beta}_{DB-site-super}\right] = \frac{1}{(J-1)J}\sum_{j=1}^{J}\left(\hat{B}_j - \hat{\beta}\right)^2 = \frac{1}{J}\widehat{\text{var}}(\hat{B}_j). \tag{7}$$

To further motivate this overall uncertainty estimate, consider that for a single, randomly chosen site we would have

$$\text{var}\hat{B}_j = \mathbb{E}\left[\text{var}\hat{B}_j|B_j\right] + \text{var}\mathbb{E}\left[\hat{B}_j|B_j\right] = \mathbb{E}\left[\widehat{SE}^2\hat{B}_j\right] + \sigma_\beta^2.$$

The first term is simply the expected within-site standard error across sites, the second is the cross-site variation. The cross-site variation turns into additional uncertainty in estimating the overall average effect.

Remark. The attentive reader will note that Schochet's original theory for the RCT-YES software proposes the estimate of the asymptotic variance of

$$\widehat{SE}^2\left[\hat{\beta}_{DB-w-super}\right] = \frac{1}{(J-1)J\bar{w}^2}\sum_{j=1}^{J}\left(w_j\hat{B}_j - \bar{w}\hat{\beta}\right)^2,$$

where we use individual weights ($w_j = N_j/N$) or site weights ($w_j = 1/J$). The $\bar{w}$ is the average site weight of $\bar{w} = W/J$.

The weights inside the square term can give very unstable and large variance estimates, however. Equation 6 stabelizes this by moving the weights outside of the sum (this improvement was originally suggested by Schochet in personal correspondence, April 11, 2018, and has since been incorporated into the RCT-Yes software).

3.3 Adjusting for covariates

To adjust for covariates, we use the approach presented in **?**. Getting covariate-adjusted impact estimates and SEs involves a series of steps that we here present. In both cases the covariate adjustment approach for obtaining SEs is parallel to the original approach, once we have our adjusted estimates.

Obtaining adjusted impact estimates. Consider that we have P individual-level covariates that we wish to adjust for. We cannot adjust for site-level covariates (as their impact would simply get differenced out).

1. Center the covariate X variables around their grand means.
2. Fit a fixed effect linear regression of

$$Y_{ij} = \alpha_j + B_j T_{ij} + \sum_{p=1}^{P} \gamma_p X_{pij} + \epsilon_{ij}$$

   This regression gives individual adjusted block impact estimates of $\hat{B}_j$.

3. Calculate the adjusted overall impact estimate by taking the person-weighted or site-weighted average of the individual $\hat{B}_j$ (see Equation 3 and Equation 4) using our adjusted $\hat{B}_j$.

Finite population adjusted SEs. To obtain finite-population standard errors, do the following:

1. Calculate the MSE of the predictions within each block in the treatment and control groups. The more predictive our initial model is, the smaller our residuals will be which will lead to smaller standard errors.

   The estimated MSE is (for the treatment side):

$$MSE_{1j} = \frac{1}{\frac{N-P}{N} N_{1j} - 1} \sum_{i=1}^{N_j} (Y_{ij} - \hat{Y}_i j)^2.$$

   The control side is analogous. Note the adjustment for the number of covariates P with the scaling factor of $(N - P)/N$ to the sample size of N_{1j}. Without covariate adjustment we would have $N_{1j} - 1$.

2. Calculate the standard error of the individual block estimates as

$$\widehat{SE}^2 \left[\hat{B}_j \right] = \frac{MSE_{1j}}{N_1} + \frac{MSE_{0j}}{N_0}.$$

   This is simply Neyman's formula on the residuals. There can be a correction term here to handle the correlation of potential outcomes to get a slightly tighter bound. We do not do this here.

3. Calculate the final overall variance as a weighted average (with weights w_j^2 (see Equation 5) of the estimated standard errors, using the $\widehat{SE}[\hat{B}_j]$ from the prior step.

Superpopulation adjusted SEs. For superpopulation standard errors, we have to look at how variable our adjusted impact estimates are, as before.

We could simply plug our $\hat{B}_j$ into Equation 6, or we can do a further adjustment as follows:

1. Calculate the mean value of the covariates within each block. This gives $\bar{X}_{1j}, \bar{X}_{2j}$, etc.

2. Calculate the overall weighted mean of means, where, for covariate p, we weight the $\bar{X}_{pj}$ by our site weights (determined by the individual vs. site averaging choice above). I.e., calculate

$$\bar{X}_p = \frac{1}{W} \sum_{j=1}^{J} w_j \bar{X}_{pj}$$

   with $W = \sum w_j$.

3. Center all of our site-averaged covariates by these mean-of-means, i.e., calculate $\tilde{X}_{pj} = \bar{X}_{pj} - \bar{X}_j$.

4. Regress our individual, adjusted, site level impacts onto an intercept and these centered covariates, weighting by our site weights. This gives our working model of

$$\hat{B}_j = \gamma_0 + \sum_1^P \gamma_p \tilde{X}_{pj} + \xi_j$$

The γ_0 will be the same as our estimated ATE from above since we have carefully centered everything while taking account of the weights.

5. Finally take the predicted values from the above model to get adjusted-adjusted site specific impact estimates

$$\tilde{B}_j = \hat{\gamma}_0 + \sum_1^P \hat{\gamma}_p \tilde{X}_{pj}$$

6. Plug these doubly adjusted values into a modified version of Equation 6 to get the final SE. The modification is a degrees of freedom correction, which gives the following:

$$\widehat{SE}^2 \left[\hat{\beta}_{DB-w-super}\right] = \frac{1}{(J-P-1)J\bar{w}^2} \sum_{j=1}^J w_j^2 \left(\hat{B}_j - \hat{\beta}\right)^2,$$

4 Regression (Fixed effect, no interactions)

The regression estimators without interaction terms have a variety of special considerations which we unpack in the following.

4.1 Simple fixed effects with no interactions

This is the most basic ordinary least squares (OLS) method we consider. We give some notation to connect to other regression estimators discussed below. Index individuals with ij, with $j = 1, \ldots, J$ indexing the site of the individual and $i = 1, \ldots, N_j$ indexing the individual within site. Define J site dummies $S_{k,ij}$, $k = 1, \ldots, J$ with $S_{k,ij} = 1$ if individual ij is in site k; in other words, $S_{k,ij} = \mathbf{1}_{\{j=k\}}$. In the case of multiple randomization blocks in a site, we would consider the blocks as our "sites" and have a dummy for each block. Fixed effect regression is then the following model with $J + 1$ coefficients:

$$Y_{ij} = \sum_{k=1}^J \alpha_k S_{k,ij} + \beta T_{ij} + \epsilon_{ij}. \tag{8}$$

We exclude an overall intercept to have individual fixed effects for each site and no reference site. One advantage of the fixed effects model compared to design-based is the single impact

parameter reduces the degrees of freedom used by the model, which could improve asymptotic inference in smaller studies.

We discuss estimation using the matrix formulation of OLS. Let X be the $N \times (J+1)$ matrix of J columns of the J site dummy variables and a last column of the treatment assignment vector Z across the $N = \sum_j N_j$ individuals. To be explicit, X is effectively J matrices, one for each site, stacked together. Each row of X can be indexed by ij. Each row of X, X_{ij} is a vector in R^{J+1} with J site dummies, with site dummy $k = j$ set to 1 and the rest set to 0, followed by a single indicator for treatment.

The classic fixed effects regression gives $J+1$ regression coefficients, with the $J+1^{\text{st}}$ coefficient being our ATE estimate. The full vector of estimates is

$$\hat{\theta} = (X'X)^{-1}X'Y.$$

Then $\hat{\beta}_{FE} = \hat{\theta}_{J+1}$. Traditional OLS then assumes homoskedasticity and gives a variance-covariance matrix on our coefficients of

$$\widehat{\Sigma} = (X'X)^{-1}\hat{\sigma}^2$$

where the $\hat{\sigma}^2$ is the estimated residual variance (essentially the average of the squared residuals, adjusted for degrees of freedom):

$$\hat{\sigma}^2 = \frac{1}{N-J-1}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\hat{u}_{ij}^2,$$

with residuals $\hat{u}_{ij} = Y_{ij} - \hat{Y}_{ij}$.

Our standard error for the ATE is the bottom right corner element of our $\hat{\Sigma}$:

$$\widehat{SE}\left[\hat{\beta}_{FE}\right] = \sqrt{\widehat{\Sigma}_{J+1,J+1}}$$

If all sites treat the same proportion of units, this will reduce to $\hat{\sigma}^2/\left(Np(1-p)\right)$.

4.2 Why fixed effects is precision weighted

To see that the fixed effect regression is precision weighted, we can directly solve the least squares optimization. Using Equation 8, we have the loss function of

$$\ell(\beta) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij} - \alpha_j - \tau T_{ij}\right)^2.$$

Our estimates will be the minimizer of this loss.

To find the minimum, we take the derivative with respect to an α_j, getting

$$\frac{d}{d\alpha_j}\ell(\beta) = -2\sum_{i=1}^{n_j}(Y_{ij} - \alpha_j - \tau T_{ij})$$

If we set this to 0 and solve for α_j we get

$$\alpha_j = \bar{Y}_j - p_j\tau.$$

Now take the derivative with respect to τ. We will be left with only those terms with $Z_{ij} = 1$:

$$\frac{d}{d\tau}\ell(\beta) = -2\sum_{j=1}^{J}\sum_{i=1}^{n_j} Z_{ij}\left(Y_{ij} - \alpha_j - \tau\right)$$

$$-2\sum_{j=1}^{J} n_{1j}\left(\bar{Y}_{1j} - \alpha_j - \tau\right)$$

If we set the above to 0, and then plug in our expressions for the α_j, we get

$$\begin{aligned}
0 &= \sum_{j=1}^{J} n_{1j}\left(\bar{Y}_{1j} - \alpha_j - \tau\right) \\
&= \sum_{j=1}^{J} n_{1j}\left(\bar{Y}_{1j} - \bar{Y}_j + p_j\tau - \tau\right) \\
&= \sum_{j=1}^{J} n_{1j}\left(\bar{Y}_{1j} - \bar{Y}_j\right) + \tau\sum_{j=1}^{J} n_{1j}\left(p_j - 1\right) \\
&= \sum_{j=1}^{J} n_{1j}(1-p_j)\left(\bar{Y}_{1j} - \bar{Y}_{0j}\right) - \tau\sum_{j=1}^{J} n_{1j}\left(1 - p_j\right).
\end{aligned}$$

The last expression comes from $\bar{Y}_j = p_j\bar{Y}_{1j} + (1-p_j)\bar{Y}_{0j}$ and then grouping. The final step is to move the expression with τ to the other side and divide by the normalizing constant. But we see our estimate is the weighted difference in means, with weights of $n_{1j}(1-p_j) = n_j p_j(1-p_j)$.

4.3 Fixed effects, Huber-White SE

The ATE estimate is the same as OLS, above. The standard error differs, however. Huber-White standard errors seek to avoid the homoskedasticity assumption, giving the classic sandwich formula of

$$\widehat{SE}_{HW}\left[\hat{\beta}_{OLS}\right] = \frac{1}{N}\left(\frac{1}{N}X'X\right)^{-1}\widehat{M}\left(\frac{1}{N}X'X\right)^{-1}$$

with a $(J+1) \times (J+1)$ "meat" matrix of

$$\widehat{M} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} X_{ij} X_{ij}' \hat{u}_{ij}^2$$

where the $\hat{u}_{ij}$ are again the residuals from the regression.

Note how if we could assume the $\hat{u}_{ij}^2$ were all independent of the X_{ij} we can (under an expectation of random variables argument) in effect separate the sum, getting

$$\frac{1}{N} \sum_{i,j} X_{ij} X_{ij}' \hat{u}_{ij}^2 = \frac{1}{N} \left(\sum_{i,j} X_{ij} X_{ij}' \right) \left(\sum_{i,j} \hat{u}_{ij}^2 \right) = \frac{1}{N} (X'X) \sum_{i,j} \hat{u}_{ij}^2$$

which cancels one of the $(X'X)^{-1}$ terms, giving our homoskedastic formula above, up to the dividing by N rather than $N - J - 1$ to correct for degrees of freedom. This suggests we should not have the $1/N$ factor at the front, but something other. And indeed this shrinkage is often corrected via further adjustments which give the "HC1," "HC2," etc., versions of heteroskedastic robust standard error estimators. See classic texts or standard documentation for statistical software for further explanation of these corrected and adjusted heteroskedastic robust estimators.

4.4 Fixed effects, Cluster Robust SE

Cluster Robust standard errors come from a sandwich estimator with different "meat." We will obtain the final estimated SE by extracting the bottom-right entry from this matrix, as before. The difference is in how the meat is calculated.

In this case, our meat will be an average over our sites, not individuals. We first obtain vectors $\hat{u}_j = (\hat{u}_{j1}, \ldots, \hat{u}_{jN_j}) \in R^{N_j}$ of the individual residuals for each site. We use these to estimate site-specific covariance matrices which are then averaged to get our overall $\hat{\Sigma}$ as we did with the Huber-White case above:

$$\hat{M}_{cluster} = \frac{1}{N} \sum_{j=1}^{J} X_j' \hat{u}_j \hat{u}_j' X_j.$$

As a check, note the X_j are $N_j \times (J+1)$ matrices, the $\hat{u}_j \hat{u}_j'$ a $N_j \times N_j$ matrix, and thus the terms in the sum all end up being $(J+1) \times (J+1)$ matrices, which we are summing. We divide by N since each inner matrix is collapsing over N_j observations (setting aside correlation issues). To see this, imagine we replaced $\hat{u}_j \hat{u}_j'$ with $diag(\hat{u}_{j1}^2, \ldots, \hat{u}_{jN_j}^2)$. This would give our Huber-White formula above.

These clustered standard errors capture how residuals are correlated within site, averaged across the number of sites. Thus, if there is a large degree of cross-site variation, the

residuals of the treated units will tend to be more correlated, and the overall average standard errors will increase. For a technical discussion of this approach within the framework of experimental design, see **?**. This discussion explains why this estimator is targeting the superpopulation estimand; it is taking the clusters (here sites) themselves as a sample from a larger population.

Just as with Huber-White standard errors, there are again degree of freedom issues to correct for, especially when there are few sites. These issues are particularly important here as they are driven by the number of sites, not individuals. Just as with Huber-White, there are several methods for correcting the estimated standard errors to account for these concerns. In particular, the usual $CR0$ clustered variance estimator uses $1/J$ to estimate the variance, $CR1$ uses $1/(J-1)$, and the bias-reduced linearization estimator ($CR2$) uses an unbiased estimator for a weighted sample variance, which reduces to $CR1$ when the sites are equally weighted.

When J is small (which is frequently the case) these standard errors still tend to be downward biased due to not correcting for the degrees of freedom lost in the estimation process. To specifically target these small-sample concerns, **?** propose an extension of the bias-reduced linearization correction to cluster robust standard errors (**?**) that allows for fixed effects. This approach is implemented in the `clubSandwich` R package; for multisite trials, it can be directly implemented as described in the blog post cited in the footnote.[1] One can also use a variant of the so-called "wild bootstrap" to get improved inference (**?**). All of this being said, the real improvement of the more recent cluster robust methods are in their improved degrees of freedom calculations. In particular, the Satterthwaite estimated degrees of freedom gives a better value for generating appropriately long confidence intervals and reference distributions for the observed t statistics.

4.5 Fixed effects with weighting

The idea behind using weighted fixed effects regression is to undo the bias from the classic fixed effect regression by weighting the individual units so the target estimand corresponds to either the person-average or site-average impact. In particular, with weighted fixed effect regression we regress outcome onto a treatment indicator and the J site dummies, as with the simple fixed effect OLS above, but now we weight each observation by ω_{ij}, where

$$\omega_{ij} = T_{ij}\frac{p}{p_j} + (1 - T_{ij})\frac{1-p}{1-p_j}.$$

All units in a site with the same treatment have the same weight. Given this we can, for a unit with treatment z in site j give the weight of

$$\omega_{zj} = \frac{N_z}{N}\frac{N_j}{N_{jz}}.$$

[1] See `https://www.jepusto.com/handmade-clubsandwich/`.

Under these weights, for each site j we have a total weight, using the above weights, of

$$\Omega_j = N_{j1}\frac{N_1 N_j}{N N_{j1}} + N_{j0}\frac{N_0 N_j}{N_{j0}} = \frac{1}{N}\left(N_1 N_j + N_0 N_j\right) = \frac{N_j}{N}\left(N_1 + N_0\right) = N_j$$

i.e., the original total number of units.

The proportion of weighted units that are treated in site j, however, is now

$$p_j = \frac{\text{mass of treated units}}{\text{total mass of units}} = N_{j1}\frac{N_1 N_j}{N N_{j1}}/N_j = \frac{N_1}{N}$$

the *overall* proportion of units treated.

When we weight units by these weights when determining our best fit line, we are mimicking having the same proportion of treated units in each site, while preserving the original sites sizes. This means our overall treatment impact estimate is going to be a site-size weighted average of the site-average estimated treatment impacts, as desired. We will therefore exactly match the person-weighted design based estimator. In particular, when running our weighted FE regression, the weights modify our "precision weighted average" of $p_j(1-p_j)N_j$ to all share the same $p(1-p)$. This means all of our precisions are proportional to site size, and we obtain our unbiased result.

Standard Errors for the weighted regression. While the weighted FE regression will give the same point estimates as corresponding design-based estimators, the method for calculating the standard errors will be different. Unfortunately when it comes to standard errors, the "weighted" for weighted linear regression is not quite fully specified. In particular, there are two distinct views of what the weights are. These views are most easily understood from the classic modeling perspective where a unit has an expected value (from the regression) and a residual that is an additional random component (such as from measurement error). The first view assumes the weights index heteroskedsaticity directly. In particular, a unit with a larger weight is more precisely estimated than a unit with a low weight; the implicit assumption is that the variance of the residual ϵ_{ij} for unit ij is inversely proportional to weight ω_{ij}. The second view takes the weights as a representation of what share of the superpopulation the unit represents; this view weights different units more heavily in estimation, but does not assume that this means we have a more precise sense of the "true value" of any given unit. In other words, in the first sense a unit with weight 2 and an outcome of 5 tells us with "twice the certainty" that the true value of that unit is about 5. In the second sense, the unit is standing in for two units, but we are not relatively more certain as to what its outcome is, as compared to other units.

Traditional homoskedastic OLS standard errors that you get from incorporating weights give the uncertainty associated with the first view. We, however, are in the second: the weights are incorporated in order to change the average estimate but do not reflect differing levels of precision across our units. Therefore, for estimation we cannot use traditional OLS but instead need to use the methods from, e.g., survey sampling, which views the weights as sampling weights (which they are). Alternatively, one can use cluster robust or heteroskedastic robust standard errors, which weights the residuals in their aggregation step; this is akin to the second view.

Estimation with weights. As discussed above, you need to use a survey regression package or use robust standard errors. In particular, for SAS, use `PROC SURVEYREG`. In R, use the `srvglm` package (and do *not* simply pass weights to the `lm()` command). One can also simply pass the weights to a robust standard error calculation.

Site weighting. If we want the average ATE across sites, not individuals, we can tweak the unit weights to produce impacts averaged across sites. We can do this by scaling the original person weights by the ratio of the average site size to the individual site size of each person's site:

$$\omega_{ij}^{site} = \omega_{ij} \cdot \left[\frac{\bar{N}}{N_j} \right],$$

where $\bar{N}$ is the average site size of $\frac{1}{J} \sum_{j=1}^{J} N_j = N/J$.

Our total weight for each site is now

$$\Omega_j = N_j \cdot \frac{\bar{N}}{N_j} = \frac{N}{J}$$

so each site receives the same weight (of the average number of units in the experiment). The proportion of weighted units in each site is still going to be N_1/N since all the units in each site have the same site rescaling of weights.

This gives final weights of

$$\omega_{ij}^{site} = \left[T_{ij} \left(\frac{p}{p_j} \right) + (1 - T_{ij}) \left(\frac{1-p}{1-p_j} \right) \right] \left[\frac{\frac{N}{J}}{N_j} \right]$$

If site sizes vary a lot, then this weighting can hurt precision. However, the expected value of the treatment effect estimate is the site average effect. Just like the individual-weight weighted estimator, this weighting approach forces the weighted proportion assigned to treatment to be the same at each site and equal to the overall proportion assigned to treatment. However, now the weighted total size of each site is equal to the average site size (i.e., $N_j^{\omega - site} = \frac{N}{J}$). Due to this, the overall average impact estimate, $\hat{\beta}_{FE-weight-Site}$, will be an equally weighted average of the individual site impacts.

Connections to the design-based estimator. Inverse weighting, in terms of the point estimate, is identical to the design-based estimator. We next show this connection mathematically. Consider the following expansion of the classic randomized block trial estimate of

the ATE:

$$\begin{aligned}
\hat{\beta}_{DB-Persons} &= \sum_{j=1}^{J} \frac{N_j}{N} \hat{B}_j \\
&= \sum_{j=1}^{J} \frac{N_j}{N} \left(\frac{1}{N_{j1}} \sum_{i=1}^{N_j} T_{ij} Y_{ij} - \frac{1}{N_{j0}} \sum_{i=1}^{N_j} (1 - T_{ij}) Y_{ij} \right) \\
&= \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{N_j}{N} \frac{1}{N_{j1}} T_{ij} Y_{ij} - \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{N_j}{N} \frac{1}{N_{j0}} (1 - T_{ij}) Y_{ij} \\
&= \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{1}{p_j} T_{ij} Y_{ij} - \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{1}{1 - p_j} (1 - T_{ij}) Y_{ij} \qquad (9) \\
&= \frac{1}{N_1} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{p}{p_j} T_{ij} Y_{ij} - \frac{1}{N_0} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{1 - p}{1 - p_j} (1 - T_{ij}) Y_{ij} \qquad (10)
\end{aligned}$$

Equation 9 is the standard IPW weighting formula. Now use $p = N_1/N$ to get Equation 10, which contains the IPW weights discussed above. In particular, the above shows that we have

$$\hat{\beta}_{DB-Persons} = \frac{1}{N_1} \sum_{i,j} T_{ij} \omega_{ij} Y_{ij} - \frac{1}{N_0} \sum_{i,j} (1 - T_{ij}) \omega_{ij} Y_{ij}.$$

An incorrect weighting approach. Some natural weighting schemes can actually lead to poor inference when using survey sampling or weighted classic regression. In particular, one might propose the seemingly simpler weights of

$$\omega_{ij} = T_{ij} \frac{1}{p_j} + (1 - T_{ij}) \frac{1}{1 - p_j}$$

i.e., weights without incorporating the overall proportions of treated or control. These weights can be expressed as

$$\omega_{zj} = \frac{N_j}{N_{jz}}.$$

Our weight for each site is now

$$\Omega_j = N_{j1} \frac{N_j}{N_{j1}} + N_{j0} \frac{N_j}{N_{j0}} = 2N_j.$$

The total weight in each site is now twice the original weight.

The proportion of weighted treated units in each site is

$$p_j = \frac{\text{mass of treated units}}{\text{total mass of units}} = N_{j1} \frac{N_j}{N_{j1}} / 2N_j = \frac{N_j}{2N_j} = \frac{1}{2}.$$

This means that we will have a precision-weighted average of each site proportional to the site size, recovering our post-stratified estimate as before. However, the proportion treated in each site is now 1/2, not p, meaning the standard errors will be a different average of the residuals in the treatment and control groups. In particular, if the treated group is a small proportion of units and the treatment outcomes have a different variance, this would not take such imbalance into account when calculating the overall uncertainty. When passing these weights to classic regression or survey regression packages, this rescaling can undermine the inference, giving systematically wrong standard errors.

5 Regression (Fixed effect with interactions)

For this approach, fit a model with all site by treatment interactions. Extending the fixed effect model above, this gives the following model with $2J$ coefficients:

$$Y_{ij} = \sum_{k=1}^{J} \alpha_k S_{k,ij} + \sum_{k=1}^{J} \beta_k T_{ij} S_{k,ij} + \epsilon_{ij}$$

We do not have either an overall intercept or an overall treatment term in the above.

Using the usual OLS regression framework we can get an estimated covariance matrix $\widehat{\Sigma}$ for the (α, β) vector. The diagonal of this matrix are the individual squared standard errors for the vector of parameters. To get point estimates and uncertainty on aggregate parameters, we define contrasts c, with $c \in R^{2J}$, which define how we add up the component pieces. These give estimates of $\hat{\beta}_c = c'(\hat{\alpha}, \hat{\beta})$, with associated standard errors of

$$\widehat{SE}^2 \left[\hat{\beta}_{OLS-c}\right] = c'\widehat{\Sigma}c. \tag{11}$$

Because the site dummies are, by design, independent from each other, the only non-zero diagonal terms of $X'X$ and $X'X^{-1}$ will be at $(j, J+j)$ and $(J+j, j)$ for $j = 1, \ldots, J$. As $\widehat{\Sigma} = (X'X)^{-1}\hat{\sigma}^2$, $\widehat{\Sigma}$ will be diagonal in the lower right quarter, giving a simple expression for the standard error of

$$\widehat{SE}^2 \left[\hat{\beta}_{OLS-c}\right] = \sum_{j=1}^{J} c_{J+j}^2 \widehat{\Sigma}_{J+j,J+j}. \tag{12}$$

for any c with the initial J terms equaling zero.

This can be further simplified. The inversion of $X'X$ can be directly solved, giving, for $J = 1, \ldots, J$:

$$\begin{aligned}
(X'X)^{-1}_{j,j} &= \frac{1}{N_j - N_{1j}} \\
(X'X)^{-1}_{J+j,J+j} &= \frac{N}{N_{1j}\left(N_j - N_{1j}\right)} = \frac{1}{p_j\left(1-p_j\right)N_j} \\
(X'X)^{-1}_{j,J+j} = (X'X)^{-1}_{J+j,j} &= -\frac{1}{N_j - N_{1j}}
\end{aligned}$$

All other terms are 0. To see this, consider that if the columns and rows of $X'X$ are reordered to keep each site as pairs, one would get a block diagonal matrix of 2×2 matrices. Each matrix will have N_j in the top-left, and the total number of treated, N_{j1} in the other three spots. The above shows that, for any c with the initial J terms equaling zero, we have

$$\widehat{SE}^2\left[\hat{\beta}_{OLS-c}\right] = \hat{\sigma}^2 \sum_{j=1}^{J} c_{J+j}^2 \frac{1}{p_j\left(1-p_j\right)N_j}. \tag{13}$$

Using Huber-White one could get alternate estimates of the $\widehat{\Sigma}$ and then use Equation 11 to get the standard error. We do not further explore these issues here.

Interaction models with individual weighting. To calculate the overall ATE let $c = (0, \ldots, 0, N_1, \ldots, N_J)/N$ (with J leading zeros) which gives

$$\hat{\beta}_{FE-inter-person} = c'(\hat{\alpha}, \hat{\beta}) = \sum_{j=1}^{J} \frac{N_j}{N}\hat{\beta}_j$$

Use Equation 12 to get the corresponding standard error estimate,

$$\widehat{SE}^2\left[\hat{\beta}_{OLS-int-persons}\right] = \sum_{j=1}^{J} \frac{N_j^2}{N^2}\widehat{SE}^2\left[\widehat{\beta}_j\right].$$

This is the equation easiest to use with the OLS output one would get from a statistical program. That being said, this could in turn be simplified to

$$\widehat{SE}^2\left[\hat{\beta}_{OLS-int-persons}\right] = \sum_{j=1}^{J} \frac{N_j^2}{N^2}\frac{1}{p_j\left(1-p_j\right)N_j}\hat{\sigma}^2 = \sum_{j=1}^{J} \frac{N_j}{N^2 p_j\left(1-p_j\right)}\hat{\sigma}^2$$

Interaction models with site weighting. Let $c = (0, \ldots, 0, 1/J, \ldots, 1/J$ and use the above. This gives

$$\widehat{SE}^2\left[\hat{\beta}_{OLS-int-sites}\right] = \sum_{j=1}^{J} \frac{1}{J^2}\widehat{SE}^2\left[\widehat{\beta}_j\right]$$

which does not really simplify, but can be written as

$$\widehat{SE}^2\left[\hat{\beta}_{OLS-int-sites}\right] = \sum_{j=1}^{J} \frac{1}{J^2}\frac{1}{p_j\left(1-p_j\right)N_j}\hat{\sigma}^2.$$

6 Random effects (multilevel modeling)

Multilevel modeling allows for having a random treatment impact for each site, but tying these impacts together with a common distribution (this is known as partial pooling). The point estimates and standard errors from a multilevel model come from maximum likelihood estimation (or restricted maximum likelihood). That being said, **?** discuss the adaptive nature of these estimators by exploring what the point estimate of the ATE would be under known cross-site treatment heterogeneity. They focus on the FIRC. We first give the model for the classic RIRC and then turn to FIRC. We finally briefly discuss another multilevel model, the Random Intercept, Constant Coefficient (RICC) model, which does not allow for treatment variation and is thus actually a variant of the fixed effect regression discussed above, despite its multilevel modeling aspects.

Regardless, the usual estimates of the standard errors come from the (potentially restricted) maximum likelihood estimates. In particular, maximum likelihood estimates are asymptotically normal with variance defined by the Fisher information. This gets empirically estimated and plugged in as part of the model fitting process.

6.1 Random Intercept, Random Treatment Coefficient (RIRC)

The RIRC model is

$$Y_{ij} = \alpha_j + B_j T_{ij} + \epsilon_{ij}$$

with

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\tau \\ & \tau^2 \end{pmatrix} \right]$$

and $\epsilon_{ij} \sim N(0, \sigma^2)$. Here α is the overall average control outcome across sites and β is the average treatment effect across sites.

As discussed in the main paper, this model heavily suggests an infinite superpopulation model, but one can also interpret this approach from a Bayesian perspective where we are "borrowing strength" or "partially pooling" coefficients across sites. In particular, the fully interacted OLS model has each site's impact estimate dependent only on that site's data. This is the unpooled estimate. The OLS method without interactions, by comparison, fully pools the treatment impact coefficient, assuming it is constant across sites. Multilevel modeling is in the middle; it partially pools, shrinking each site estimate towards the overall mean. This can give superior properties when looking at individual site impacts. It also could arguably stabilize the overall treatment estimate. See **?** for further discussion in textbook form.

Uncertainty estimates from this approach are superpopulation uncertainty estimates. This means that even if the actual finite population properties of this estimator are strong, we might not be able to measure the benefit from them without any corresponding finite population estimates of the uncertainty. The superpopulation estimates will be overly conservative. Developing finite population uncertainty measures for these models is an area for future work.

6.2 Fixed Intercept, Random Treatment coefficient (FIRC)

The FIRC model lets α_j be a fixed effect (so we do not put a model on it). This is equivalent to forcing the σ_α^2 term from the RIRC model to ∞ (when viewing it as a prior). This model has the same structural form as the RIRC model, but only has a random effect for the B_j:

$$B_j \sim N(\beta, \tau^2)$$

where σ_β^2 is the cross-site treatment effect variation.

? show that, if σ_β^2 were known and the residual variances in each site were known, then the estimate $\hat{\beta}$ is a weighted average of the site average impacts (see Equation 2) with weights of

$$w_j = \left(\sigma_\beta^2 + V_j\right)^{-1} = \left[\tau^2 + \frac{\sigma^2}{N_j p_j (1 - p_j)}\right]^{-1}$$

The second equality comes from the homoskedasticity assumption in the residuals across site. This equation shows the adaptability of FIRC: as cross-site treatment effect variation increases, it will dominate the weight, making all site weights converge to $1/\sigma_\beta^2$, which will equally weight sites. If, however, there is no site variation, we end up recovering the precision-weighting of fixed effect OLS. In practice σ_β^2 is not known, and FIRC essentially plugs in an estimate to get the overall $\hat{\beta}$.

If we assume a constant site size and constant proportion of units treated within each site, the standard error of this estimator can be expressed as

$$SE\left[\hat{\beta}\right] = \left[\frac{1}{J}\tau^2 + \frac{\sigma^2}{Np(1-p)}\right]^{1/2}$$

For discussion see **?**, where the present the above for the purposes of calculating minimum detectable effect sizes. For varying site size and proportion treated, this can be viewed as an approximation.

6.3 Random Intercept, Constant Coefficient (RICC)

This model has only a random intercept term and a constant treatment parameter shared across all sites. This makes it analogous to the classic fixed effect regression discussed above. In particular, the treatment impact will essentially be a precision weighted estimate; partially pooling the fixed effects only has a modest impact on the treatment estimate, as compared to the fixed effects regression approach.

Even given the random intercept, uncertainty estimates from this approach are surprisingly finite-sample; one way to think about this is the treatment estimate does not involve any of the random effects, so the hypothetical sampling of site intercepts has no impact on the point estimate. Under the assumption that the treatment is constant, the associated standard errors have no ability to incorporate the sampling uncertainty of any potential cross site variation.

7 Further considerations in estimation

There are several considerations that cut across the different estimation strategies that we now discuss. First, all modeling strategies have a residual variance component that itself is a model for how individual units vary around their expected site means within a site. There are several choices about how to model this residual variance that are related. Second, in many designs there may be multiple randomization blocks within any given site; for example, if there are multiple cohorts across a series of years in a school-lottery analysis, the site would be the school and the randomization blocks would be the cohorts. This introduces further complications in interpreting both the estimands and implementing the estimators. It also makes the superpopulation sampling assumption less clear.

7.1 The Residual Variance Model

For linear regression (without robust standard errors) and multilevel modeling approaches, uncertainty estimation is tied to what kind of structure one is willing to place on the residual variation, that is the variation of individual units around their site or overall means. Classic linear regression, for example, assumes homoscedasticity, i.e., that the amount of variation is the same within treatment arms and across sites. Especially when one believes there is treatment effect variation, this can be overly restrictive (consider, for example, that a simple additive model of heterogenous treatment would generally induce greater variation in the treatment group than control). We might also expect variation within each site to potentially be different than other sites.

The most extreme is to allow each site and treatment arm to have its own residual variance. Unfortunately this could result in a large number of very unstable variance estimates (two per site), especially if some sites are small. In the fully interacted linear model case, these uncertainty estimates will then be averaged together to form the standard error for the ATE. Even without interaction terms, the individual residual uncertainties will still all play a role in the overall estimated standard errors. The finite-sample design-based procedures explicitly work this way; the robust standard error procedures put no model on the residuals other than independence assumptions, and so are in effect working this way as well.

For the FIRC model, these plethora of individual variance terms can create serious problems for estimating the amount of cross site variation, as discussed in **?**. A compromise approach, also presented in **?**, is to allow a different residual variance term for the treated units and control units. In particular, the model would be, extending the fixed effect model of Equation 8,

$$\epsilon_{ij} \sim N\left(0, \sigma^2_{T_{ij}}\right)$$

with σ^2_0 and σ^2_1 being two distinct residual variance parameters.

Regardless, the structure put on the residuals is a separate choice from the form of the least squares model. Even the fully interacted model with different treatment impacts per site could be fit under the classic homoscedasticity assumption. In this case while you will get unbiased estimates for the treatment effect for each site, the overall uncertainty estimate will depend on, in part, this homoscedasticity assumption.

Finally, one can avoid explicit modeling by using heteroskedastic robust standard errors. These model covariances using the residuals of a fit model in a non-parameterized way. Cluster robust standard errors go a step further, allowing for the units within a site to be arbitrarily correlated; these fundamentally take the sites themselves as the units of analysis.

7.2 Randomization blocks within sites

In most of the real-world experiments we examined in our empirical studies, randomization of individuals occurred within blocks, not sites. In particular, each site (e.g., a school), would contain more than one block if randomization was done on features such as cohort or grade. Here, we discuss the nuances of dealing with this complication.

When there are multiple randomization blocks within the site, there are two general approaches one might take. The first is to simply consider the individual randomization blocks as sites in their own right, and proceed. The second is to aggregate or restrict the block estimates to account for their nesting within site; this is the approach we took in this work. We next discuss how randomization blocks within sites can impact the three modeling approaches below.

Design based approaches. Handling randomization blocks for the design based approaches is relatively straightforward in the finite-population context. For person-weighting, simply treat the randomization blocks as sites, and weight by their size as before. For site-weighting, one would weight the individual randomization blocks by their fractional size of the corresponding site, so the weights would add up to one within each site. This corresponds to a three-step process of first obtaining the individual block level estimates, then taking a weighted average of the individual block effects within each site to obtain the site level estimates, and then finally averaging the sites to obtain the overall estimate.

We need to add an index to make this more precise. Let r index the randomization block and, as before, j index the site. Then, if we have randomization block estimates of $\hat{b}_{rj}$, and n_{rj} units in randomization block r of site j we can write our site-level estimates of ATE as (letting R_j denote the number of blocks in site j):

$$\hat{B}_j = \sum_{r=1}^{R_j} \frac{n_{rj}}{N_j} \hat{b}_{rj}.$$

Our overall site-average estimate is then

$$\hat{\beta} = \frac{1}{J}\sum_{j=1}^{J}\hat{B}_j = \sum_{j=1}^{J}\sum_{r=1}^{R_j}\frac{1}{J}\frac{n_{rj}}{N_{rj}}\hat{b}_r.$$

The above shows that we are weighting the randomization blocks by $n_{rj}/(JN_{rj})$. Our finite-sample design-based estimators can simply use these weights in the impact and standard error estimator formulae.

For the standard errors for superpopulation inference, things are not quite as straightforward because the posited sampling scheme is no longer clear (in particular, we cannot think of the blocks as being a random sample of a superpopulation of blocks, and the design based estimators do not account for varying chances of being treated for different blocks within a site). To generate standard errors we therefore consider the blocks within site as fixed, and do the following: first aggregate the block estimates $\hat{b}_{rj}$ by site to get estimated site impacts $\hat{B}_j$, as described above. Second, plug the site estimates into the superpopulation formula that looks at cross site variation to assess overall variation. This approach assumes that the sites are sampled from a superpopulation of sites, but that each site comes with its blocks and treatment proportions across blocks all predefined and fixed. In other words, we are sampling a series of small, finite-sample blocked experiments. We leave a more technical treatment of this approach to future work.

Regression approaches. For the linear regression approaches, the precision-weighted average from the simple fixed effect model, if using block fixed effects, will be a precision-weighted average of the blocks, not sites. Researchers should take that into account when interpreting their resulting impact estimate, but essentially the story is the same: the estimator is a biased person-weighted estimator, but the bias is likely not too large. If using some form of cluster robust standard errors, the clustering needs to be at the site level, not block, to allow for dependencies across the blocks due to shared site characteristics.

To implement inverse weighting approach for person-average estimates, rescale at the block level instead of site:

$$\omega_{zrj} = \frac{N_z}{N}\frac{n_{rj}}{n_{zrj}}.$$

If one is targeting the average impact across sites by including extra weights, a modicum of care is needed for handling the randomization blocks. In this case, the weights need to be fractional, as discussed in the design based case above. In the simple case of no randomization blocks, recall how the initial site-weighting is simply to rescale the individual weights by the relative size of the average randomization block to the specific randomization block ($\bar{N}/N_j$). Now we want to scale each randomization block to the portion of the average site that it represents. This entails scaling the individual weights from above by $\bar{N}/N_{rj}$ where the $\bar{N}$ and N_{rj} are the average size of the sites and the size of the site holding the randomization

block in question. I.e., for individual i who has treatment z in randomization block r in site j we have the following weight

$$\omega_{zrj}^{site} = \left(\frac{N_z}{N}\frac{n_{rj}}{n_{zrj}}\right) \cdot \left[\frac{\bar{N}}{N_{rj}}\right].$$

The n_{rj} terms get cancel, in effect, giving the original scaling.

For the interacted modeling approach, the coefficients are estimated at the randomization block level as described above. Now, however, we need to use a different vector of weights c to target the site averaged impact. In particular, we have $c_r = n_{rj}/(JN_j)$, as with the site-averaged design-based estimator above. We then plug that in, and thus our standard error estimate is then a sum across randomization blocks:

$$\widehat{SE}^2\left[\hat{\beta}_{OLS-int-site}\right] = \sum_{r=1}^{R} \frac{1}{J^2}\frac{n_{rj}^2}{N_j^2}\widehat{SE}^2\left[\widehat{\beta}_r\right].$$

Multilevel modeling approaches. For the multilevel modeling approaches some consideration needs to be made to the independence assumptions required for inference. In particular, the multiple blocks within a site are likely to be correlated, and in particular both the baseline block means and the impacts themselves could be coupled within site. Again following **?**, we advocate having a fixed effect for the randomization block, and making an assumption of homogenous impacts across the randomization blocks within site. This is easy to do with the FIRC model and translates to a random effect for the impact for the entire site that is shared by all randomization blocks within the site, and a fixed effect for each randomization block. This gives the following model:

$$Y_{ij} = \alpha_{bj} + B_j T_{ij} + \epsilon_{ij}$$

with

$$B_j \sim N(\beta, \tau^2),$$

as before.

For the RIRC model, one could have a shared random intercept as well as random impact for each site, across randomization blocks, but this will not be correct if the proportion of units treated varies across blocks within a site. In this case, one could try an unwieldy three-level model with individuals nested within blocks, and blocks nested within sites. The issue with random effects only at the block level is the dependence between blocks within site will not be accounted for, resulting in potentially overly optimistic standard error estimates. This is why we did not fit the RIRC model to the data in our emperical examples.