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

Online Supplement B: Simulation Study to Accompany Estimands, Estimators, and Estimates Paper

September 5, 2020

In this supplement we present the details of our simulation study along with additional results not covered in the main paper. We first discuss the simulation itself, and then present, in Section 2, results for the performance of the point estimates, and finally present, in Section 3, results regarding the standard error estimates.

1 Simulation Setup

In our simulations, we explore data generating processes (DGPs) spacing a range of scenarios just beyond those observed in our empirical examples. Our DGPs follow a classic multilevel modeling structure, with individuals nested in sites. We first generate sites and site sizes, and then generate the individuals within. In particular, our data generating model is based on the following:

$$Y_{ij} = \alpha_j + \beta_j Z_{ij} + \epsilon_{ij} \tag{1}$$

$$\alpha_j = \gamma_0 + u_{0j} \tag{2}$$

$$\beta_j = \gamma_1 + u_{1j} \tag{3}$$

with

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\tau \\ & \tau^2 \end{pmatrix}\right]$$

and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

Using this model, we generate data in a series of steps:

1. Generate site sample sizes for the sites.
2. Generate the site random effects with a bivariate normal distribution using the model above.

3. Generate the proportion of units treated within each site, if it varies.
4. If needed, link site average impact to site size or proportion treated, as described below.
5. Generate individuals for each site by generating N residuals according to our model.
6. Assign units to treatment.
7. Calculate observed outcomes according to the model above.

We simulate under a correctly specified multilevel model; we leave exploring how the estimators behave under nonnormality in residuals or other features to future work. We attempted to roughly mimic the structures we saw in our empirical examples, and consider a variety of scenarios indexed by several factors. These are:

- the number of sites, J, with $J = 10, 20, 40$, or 80.
- the expected total number of individuals, N, with $N = 1000, 2000, 4000$, and 8000.
- the amount of cross site treatment effect variation, τ^2, with $\tau^2 = 0, 0.1^2$, and 0.2^2.
- a four way factor of whether proportion of units treated in a site is constant, or whether it varies and is negatively correlated, positively correlated, or not correlated with average impact of the site.
- a three-level factor of whether site size is constant (not varying), whether it varies independently of impact, or whether it varies and is positively correlated with impact.[1]

We ran a factorial experiment, with different levels of J, N, and τ^2. These factors along with the two factors governing the relationship of impact with site size and proportion treated give 448 unique scenarios (we drop the correlated factor combinations with no variation in cross site impacts, as in these cases no correlation is possible). In presenting our results we reparameterize, and report in terms of the expected average site size of $\bar{N} = N/J$.

Our experimental factors allow the average treatment effect at site j, B_j, to be both correlated with the N_j (site size) and the p_j (proportion treated), meaning that the sites with the biggest impact tend to be the largest and most lopsided in terms of treatment (in the case of positive correlation with p_j) or largest and least lopsided (in the case of negative correlation).

We also have some fixed parameters and aspects of the data generation process that we calibrated, as we discuss below, using our empirical examples:

[1] For simplicity, we do not consider both positive and negative correlations. We do for proportion treated so we can have the correlations of the two factors be aligned or opposite, causing different aggregate biasing patterns.

- degree of site size variation. For those simulations where site size varies, we allow substantial variation, with the ratio of the variance of site size to the average site size being set to 0.60, which is the 75th quantile of our empirical datasets.

  When site size varies, we ensure all sites have at least 4 units. We generate site sizes from a 50-50 mixture of two uniforms, one that is uniform between 4 and the average site size ($\bar{n}$), and the second uniform from $\bar{n}$ to a maximum value M. M is set to give a desired variation in site sizes. The resulting distributions tend to have a mass of small sites with a long tail of larger sites.

- proportion of units treated. The average proportion of units treated in a site is held at 0.65, which many of our empirical examples had.

- variation in proportion of units treated. When the proportion treated varies, they vary from a bit under 50% to a bit over 90%. To do this we generate site-specific proportions from

  $$p_j \sim \text{uniform}[p - 0.75(1-p), p + 0.75(1-p)].$$

  We then adjust p_j in small sites so there are at least 2 units in treated and in control for all sites (to avoid definition concerns with some of the estimators).

- ICC. We fixed the ICC (how much of the control-side variation is explained by site) to 0.20, which is at the high end of our estimated ICCs in the empirical data.

- ATE. We fix $\beta_{SP-Sites}$ to be 0.20.

For each scenario we simulated in a nested structure so as to be able to estimate finite-sample performance. That is, we did the following 1000 times:

1. Generated a dataset using a particular DGP. This data generation is the "sampling step" for a superpopulation (SP) framework. The DGP is the infinite superpopulation. Each dataset includes, for each observation, the potential outcome under treatment or control.

2. Record the true finite-sample ATE, both person and site weighted.

3. Then, three times,[2] do a finite simulation as follows:

   (a) Randomize units to treatment and control.

   (b) Calculate the corresponding observed outcomes.

   (c) Analyze the results using our full set of methods, as listed Table 1 of the main paper, recording both the point estimate and estimated standard error for each.

[2] Having only three trials will give a poor estimate of within-dataset variability for each dataset, but the average across the 1000 datasets in a given scenario gives a reasonable estimate of expected variability.

For each set of estimates we also record the differences of the point estimates of each method to the four possible estimands (finite or superpopulation ATE weighted by site or person). (We estimate $\beta_{SP-Persons}$ for each scenario by taking the average of the true finite person-weighted ATEs across all datasets generated by that scenario's DGP. $\beta_{SP-Sites}$ is directly set as a parameter in the DGP.) These give the actual estimation errors for each estimand. We summarize these estimation errors to analyze the performance characteristics of the estimators.

Our simulation study focuses on three general questions:

1. How well the point estimates of the estimators perform *in truth*?

2. How well the estimated standard errors estimate the true standard errors?

3. How well the estimators do in terms of coverage?

The first question revolves around actual standard error, actual bias, and actual RMSE (root mean squared error) of the estimators with respect to the four estimands. The second question revolves around looking at the average and standard deviation of the *estimated* standard errors for each method, and comparing these estimates to the corresponding true SEs. The final question is a consequence of both the first and the second question, and unpacks how bias can undermine inference.

1.1 How we generate site sizes and proportion treated

For many of our simulations, when site size varies, we want to induce a relationship between site size (n_j) and/or proportion of units treated within a site (p_j) to site average impact (the β_j). We do this by ordering our random effects by a sorting index which is equal to the random impact β_j plus some additional noise. Then, if we want site impact to correlate with site size, we order our site sizes from smallest to largest; this makes site size roughly correlate with site impact, depending on the noise parameter in the prior step. If we want a site-impact by proportion treated correlation either sort these proportions in increasing or decreasing order. We control the strength of the relationship by selecting how much noise to add to the sorting index. We selected this parameter to give overall correlations between site size and site impact and/or proportion treated and site impact in line with the high end of our empirical studies.

1.2 How we calibrated our simulations

The levels of the factors discussed above were selected based on our empirical data. In particular, we targeted 65% treated to give a lopsided treatment chance, similar to many of our studies, and to allow for bias. Centering treatment at 50% would reduce sensitivity to the bias due to precision-weighting. We let the proportion vary a bit more than what we saw in the more variable of our empirical examples (the standard deviation of the proportion treated varied from around 0.10 to a max of 0.20, with the largest empirical variation being 0.16). This allows potentially greater benefits and costs for the fixed effect models that use precision weighting. (We also include scenarios with no variation in proportion treated to set aside these impacts and concerns.)

We set the ICC to 0.20, which was at the high end of the empirical estimates we had from our studies. We obtained these estimates using a simple random intercept multilevel model.

We set number of sites and site size to bracket our evaluations, after dropping a very large study with more than 10,000 participants. We did end up, due to the crossed factorial nature of our study, with some average site sizes much smaller ($\bar{N}_j = 12$, 25, and 50) than we saw empirically (with a smallest average of 64).

We calibrated the variation in site size so that the ratio of the site variance to site size squared matched the more variable of our empirical examples. This is quite a large degree of variation in sites.

The strength of the correlations of site impact to site size and proportion treated, when they are present, are controlled by a single tuning parameter. We first order true site impacts, site sizes, and proportions treated to share the same rank order, and then shuffle the site impacts to a degree controlled by this single parameter. We tuned this parameter by making sure our empirically observed correlations of the fake data were similar to the higher observed levels of our actual data (these correlations tended to be small). The superpopulation person-weighted ATEs tended to be about 0.02 higher than site-weighted for $\tau = 0.1$, and 0.04 higher for $\tau = 0.2$.

Overall, we believe our simulation scenarios tend towards the more extreme ends of what we saw empirically. This means the degree of bias and instability is also at the high end. Our simulation code is public, and exploring other parameter settings would be a good area of future exploration. We also include scenarios with no variation in site size and/or proportion treated, partially so we could see the impact of one factor when there was no impact from the other, and also to ensure that the estimation methods behaved as we expected in these simpler contexts.

2 Performance of the point estimates

Although we have many methods, several groups of these methods return the exact same point estimates. For example, the design based finite sample person weighted estimator gives the same point estimate as the inverse probability of treatment weighted estimator, the interacted linear model estimator, and the design based superpopulation person weighted estimator. We therefore, in this section, limit our attention to six estimators that represent the different point estimates.

The core six estimators we select are unbiased person-weighted (DB-Persons), unbiased site-weighted (DB-Sites), fixed effects regression (FE), FIRC, RIRC, and RICC. The first two are unbiased. The fixed effect estimator attempts to improve precision, but should generally be aligned with the unbiased person weighted estimator. The RICC estimator is essentially a version of a FE estimator. The FIRC and RIRC are adaptive, but target site-weighted estimands in principle. In initial analyses of these estimators we found that

RIRC was quite similar to FIRC, and RICC was extremely similar to FE, so we drop them from the presentation of the overall results to reduce clutter, and discuss their similarities subsequently.

2.1 Are the methods differently precise?

To examine whether the different methods are differently precise, we first examine their true standard errors. For each given scenario, each estimator has two true standard errors: the standard error across all the samples drawn from the superpopulation, and the average finite sample standard error.

The superpopulation true standard error is simply the standard deviation of the point estimates across all simulations within that scenario. This incorporates the variability due to the sampling process, as it should.

To investigate the true finite-sample standard errors, we take the variance of the three point estimates for the three randomizations within each generated dataset, and then average these over the datasets generated for a given scenario. This gives the *expected* true finite-sample variance for a given data generating process (so if the true variance depends on the data itself, such variability would be averaged out, but the variability of the shifting estimands is *not* included in these averaged standard errors). The square root of this describes how much the estimator tends to vary across randomizations within a set of given, fixed samples with shared heritage (defined by their DGP).

To ease comparisons across simulation scenarios that vary in sample size and, therefore, overall precision, we, for each scenario, standardize the true standard errors by the true (finite or superpopulation, depending on context) SE of the unbiased design-based person-weighted estimator:

$$inflation = 100\% \times \frac{SE}{SE_{DB-Persons}}.$$

This give the percent relative performance for that context. We then summarize these percentages across all the simulation contexts, reporting the middle 90% range to show the range of relative efficiencies of a given method to the baseline across simulation scenarios.

Results are on the table, below. The first pair of columns show average and spread for finite-sample inference. The second pair of columns show average and spread of the superpopulation SEs to the superpopulation SE of DB-Persons. The final pair of columns show the average and spread of the superpopulation SEs to the corresponding finite-sample SE for that same estimator; this shows the increased variability in an estimator due to the variability of the sample.

We see that, for the finite-population viewpoint, the unbiased site-weighted estimator DB-Sites is more variable than DB-Persons, with a 31% average increase. The range shows that for some scenarios they can be more than 78% larger, but that in other scenarios there is no difference. The fixed-effect estimator (and RICC), designed to have higher precision, does in fact have that, but the gains are modest (around 2% on average, with little variation across scenarios.) Interestingly, FIRC (and RIRC) can also show gains, although they can also be more variable, depending on scenario. We investigate this pattern of findings a bit further below.

In the superpopulation context, the site weighted estimator still has higher SEs. These relative numbers are muted because the overall variation of everything is inflated by the same amount due to the variability of the samples themselves. The FE estimator can now actually incur a penalty (although this is small) but FIRC is now actually relatively less variable, on average.

Finally, the final column shows that the true superpopulation standard errors for the various estimators are generally substantially larger than the corresponding finite sample viewpoint, but there is substantial variability in this. This variation comes from the different amounts of cross-site impact variation across scenarios: in some circumstances the superpopulation uncertainty is much larger due to this impact variation, and in other circumstances (i.e., when there is no impact variation so the estimands all coincide) the differences are nonexistent.

	Finite Sample		Superpopulation		Relative Performance	
Method	Mean	90% range	Mean	90% range	Mean	90% range
DB-Persons	100		100		140	(100-251)
DB-Sites	131	(100-178)	114	(100-147)	120	(100-177)
FE	98	(95-100)	99	(94-102)	143	(100-265)
FIRC	101	(96-115)	97	(90-100)	132	(100-207)

Table 1. Comparing actual SE across methods. Ranges are the middle 90% of the values across all simulation scenarios. Monte-carlo error on individual relative percents are about 3 percentage points in magnitude. Mean comparisons have standard errors of 1.4 percentage points or less, including variability across scenarios.

Underscoring the roll of cross-site variation in impacts, Figure 1 shows the relative SEs by this variation, restricting our attention to only those scenarios with variable site size, variable proportion treated, and without correlation between site size, proportion treated, and impact. We see no major trends with one interesting exception: under the finite perspective, FIRC is increasingly relatively variable when there is more cross-site impact variation, but under the superpopulation perspective, FIRC is relatively less variable as cross-site impact variation increases. This is because of the partial pooling: DB-Persons incorporates all cross-site impact variation in its estimate (weighted by site size) while FIRC will partially pool, shrinking the amount of such variation. For DB-Persons this variation is static across randomizations in a finite-population context, but FIRC will shrink differently depending on the overall estimated variation and individual site impacts. Hence it is more variable. From the superpopulation perspective, on the other hand, DB-Persons varies across datasets due to this cross-site variation combined with varying site size, and FIRC is more stable with respect to this variation as it tends towards the site-average, so the relative SE of FIRC grows relatively smaller. In other words, all the estimators are increasingly variable in the superpopulation context as site variation increases, but FIRC's pooling makes its variation grow more slowly than DB-Persons.

The apparent reduction in the relative SE of the site weighted estimator as cross site variation goes up in the superpopulation context (right hand side of plot) is a function of all the estimators getting more variable, making the relative sizes more similar.

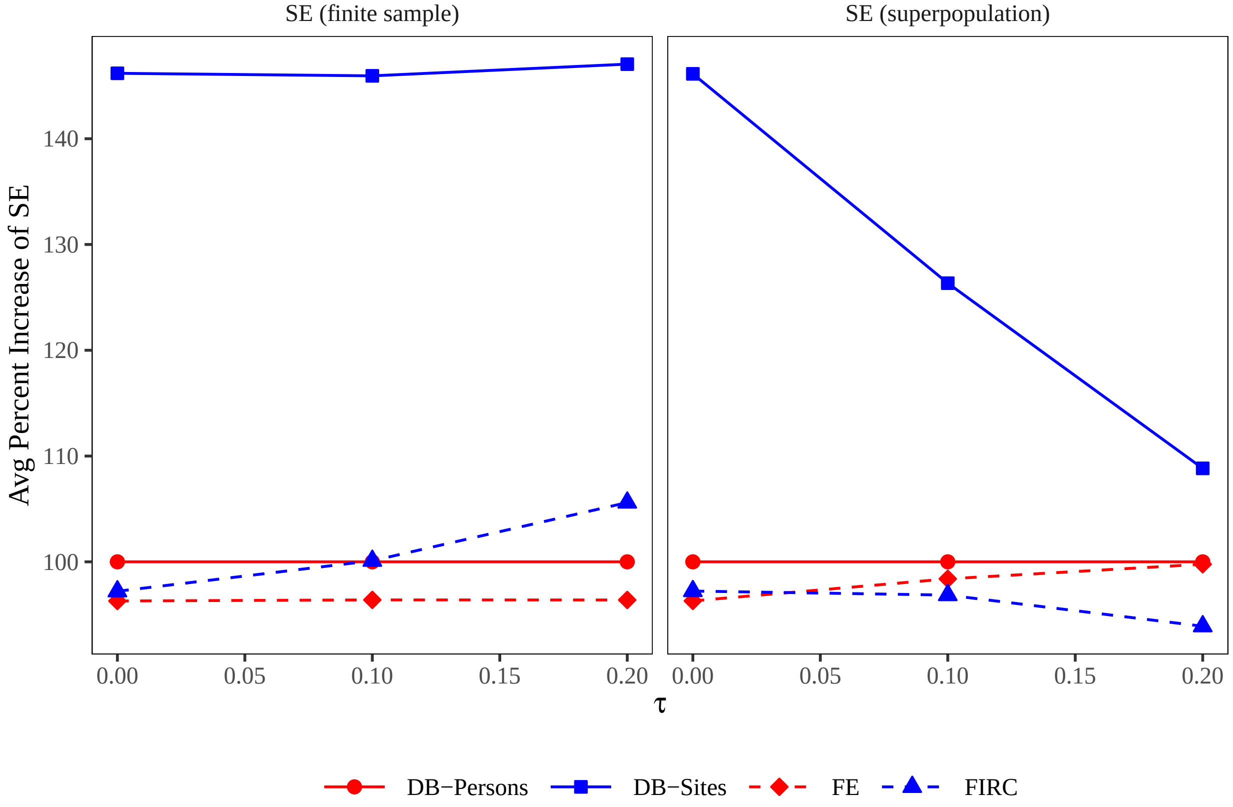

Figure 1. Percent increase of SE. The percent increase of the standard error (relative to DB-Persons) as a function of cross-site impact variation. Figure includes only those scenarios with no biases between proportion treated or site size and site impact. Left figure are finite-sample, right is superpopulation.

2.2 Are the methods different in terms of their RMSE?

In the prior section we saw some estimators indeed have smaller true SEs than others in some scenarios. But, as we have discussed, some methods are exchanging bias for increased precision. We therefore investigate the Root Mean Squared Error (RMSE) of the different estimators to take into account both bias and variance. We also show how person targeting estimators have poor RMSE for site estimands, and vice-versa, when these estimands differ.

There are four RMSEs, one for each of the four possible estimands. To measure the true RMSEs we take the square root of the average squared error:

$$RMSE_{W,F} = \left[\frac{1}{R} \sum_r \left(\hat{\beta}^{(r)} - \beta^{(r)}_{W,F} \right)^2 \right],$$

with the $F = p, s$ indicating which estimand (person-weighted or site-weighted), and the $F = f, s$ indicating which framework (finite-population or super-population). For finite-sample, the $\beta^{(r)}_{W,f}$ will be the finite-sample quantity (person or site average impact), and will vary with the sample. For the superpopulation context, $\beta_{W,s}^{(r)} = \beta_W$ across the R simulation trials.

For each scenario, we calculate the four different RMSEs corresponding to our four estimands for each of the estimators, and then compare these RMSEs to the baseline design-based unbiased estimator for that context. Numbers above 100 mean an estimator is worse (has higher RMSE) by that many percentage points above 100. Numbers below 100 mean the considered estimator tends to be closer to the estimand. We average these numbers across all scenarios and also report ranges to show variability across the scenarios. See Table 2.

To examine the circumstances when the different estimators succeed or fail, we plot, for each method, the average relative inflation for the different estimators within scenarios grouped by the correlation structure and degree of cross-site impact variation. We include plots for each of the RMSEs (although these plots are broadly similar in their patterns).

	Finite Person		Super Person		Finite Site		Super Site	
method	Mean	Range	Mean	Range	Mean	Range	Mean	Range
DB-Persons	100		100		94	(69-126)	94	(72-116)
DB-Sites	146	(100-246)	118	(100-156)	100		100	
FE	101	(96-115)	99	(96-103)	93	(68-125)	93	(72-110)
FIRC	110	(97-160)	98	(93-101)	86	(66-100)	89	(71-100)

Table 2. Ratio of the RMSE of each estimator to the RMSE for the appropriate unbiased estimator, averaged across all scenarios. Intervals are middle 90%. The four sets of columns in the table correspond to the four different estimands. Overall Monte-carlo SE for the means is 2.4 percentage points or less.

RMSE Results for person-weighted estimands. For the finite-population person-weighted estimand (first pair of columns on the Table 2), we first see that DB-Site is reliably the worst choice for the person-weighted estimand (this is unsurprising given its explicit targeting of a different estimand and observed increased variability as illustrated in the SE investigation above; it is both more variable and more biased). In fact, it can never be superior to the person-weighted, since the weighting by sites can only increase variability and there is no bias to remove.

We also see that the fixed effect estimators (FE and, by association, RICC), designed to be precision weighted, can outperform the person weighted design-based estimator, but (as shown on Figures 2 and 3) only if there is not substantial bias due to a correlation between proportion treated and impact (see top and bottom rows of the plots). Furthermore, when there is a great deal of cross-site impact variation, the variable proportion of treated, even if randomly allocated, can cause random amounts of finite sample bias (see the facet on the third row, second column of Figure 2 corresponding to independence with site size and proportion treated). Averaged across finite sample scenarios, this introduces a larger average finite-sample RMSE, which is a very surprising result.[3] From the superpopulation perspective on Figure 3, however, when p_j is independent of B_j, we do not see a penalty,

[3] For finite-sample inference, we are looking at $\mathbb{E}\left[RMSE_f\right] = \mathbb{E}\left[SE_f^2\right] + \mathbb{E}\left[bias_f^2\right]$, and since the finite sample bias depends on random fluctuations in the relationship of proportion treated to impacts, this expectation is positive, driving up the RMSE.

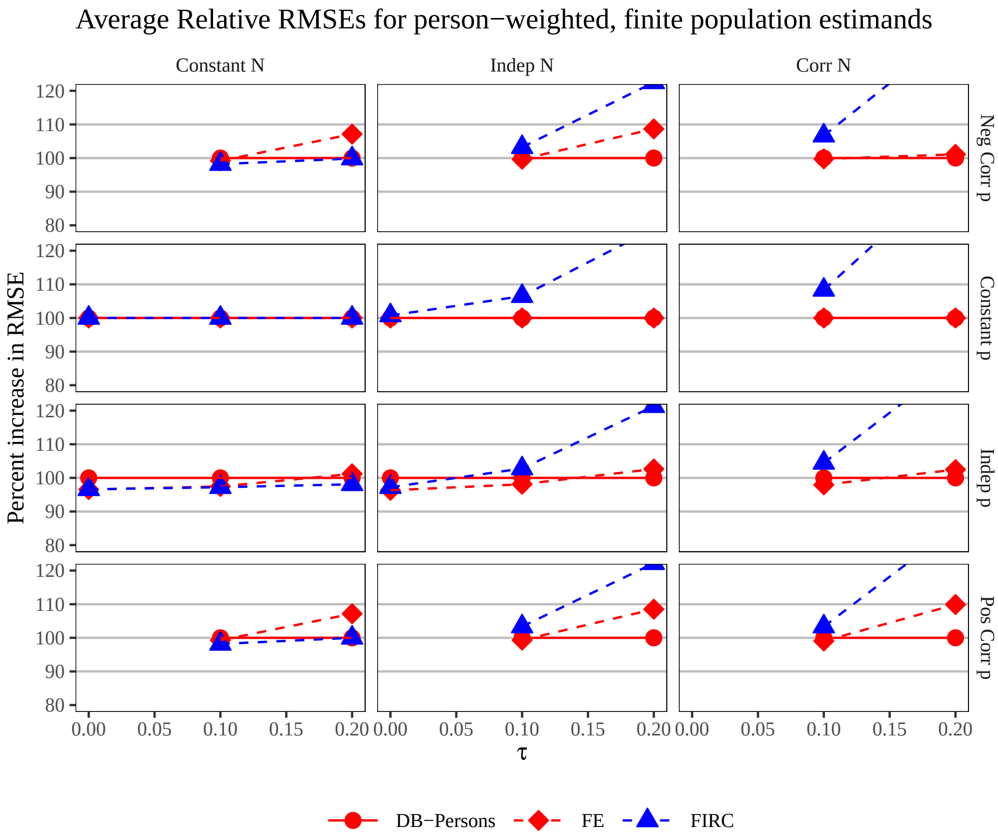

Figure 2. Relative percent increase of average finite sample RMSE to person-weighted design-based estimator. DB-Sites entirely dropped due to extremely poor performance; the plot is truncated to show detail in the 80% - 120% range. The grid of plots are for different structures in the data. From left to right we have no variation in the N_j, variation in N_j independent of B_j, and variation in N_j correlated with B_j. For the rows we have negative correlation of p_j and B_j, no variation in p_j at all, variation in p_j independent of B_j, and positive correlation of p_j and B_j.

just a loss of the potential gains (compare row 3, column 2 of this figure to corresponding facet on Figure 2).

The nominally site-weighted FIRC can also outperform the design-based estimator for the person-weighted estimands, likely due to its being a variant of the precision-weighted fixed effect estimator. For finite-sample RMSE, these gains can easily be lost when there is substantial cross-site impact variation, as the adaptive FIRC model will then tend to target the site average impact.

The superpopulation stories are roughly the same. The relative differences are shrunk due to the additional variation shared by all estimators which inflates all the RMSEs by a constant amount and makes the ratios smaller. One exception to this is, for FIRC, if there is independence of the B_j and N_j we see increasing gains in superpopulation RMSE with greater cross site variation as the site averaging targets the superpopulation average more effectively (compare these gains to the losses from the finite-sample viewpoint). When N_j is independent of B_j, the person-weighted and site-weighted estimand are the same. The improvement of FIRC is thus because DB-Persons is randomly weighting some sites more heavily than others due to site size; FIRC tends to weight them more equally. If the

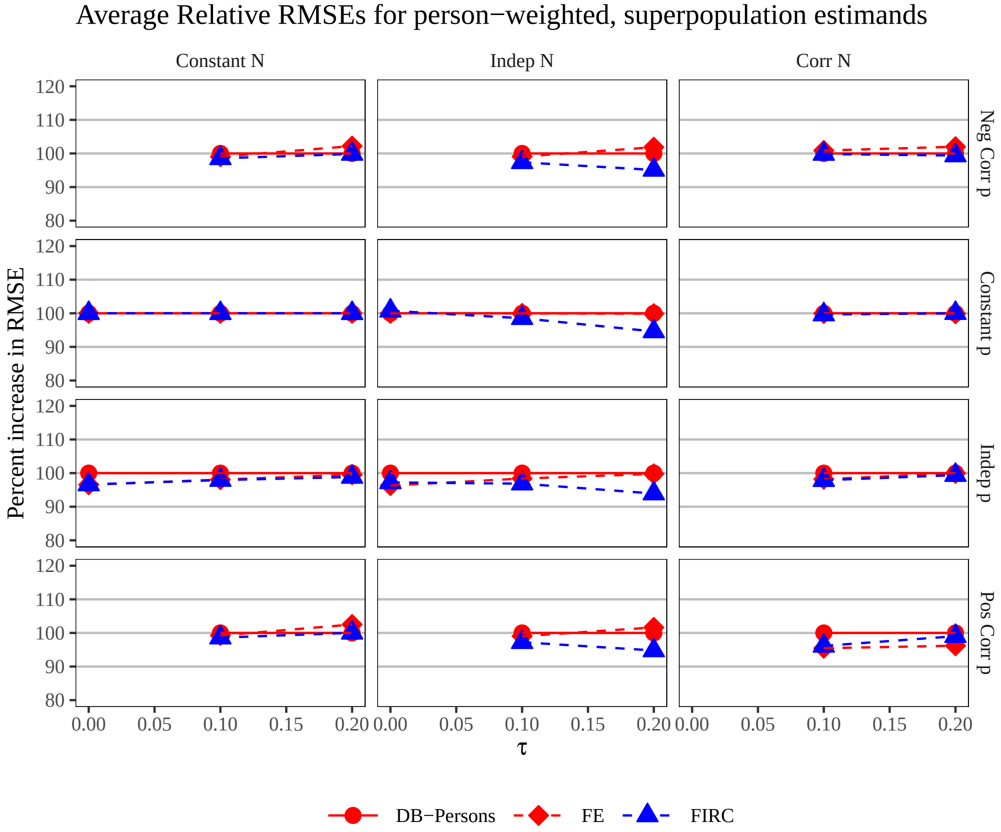

Figure 3. Relative percent increase of RMSE to superpopulation person-weighted design-based estimator. Y-axis again cropped to 120, causing DB-Sites to not be visible.

uncertainty due to cross-site impact variation is larger than the uncertainty in measuring the B_j of the sample, FIRC's strategy of more equally weighting sites leads to overall increased stability and improvement (see right hand of Figure 1).

RMSE Results for site-weighted estimands. The person weighted DB can outperform the site weighted, even when the target is the site average ATE. The DB-Sites estimator's great variability when there is substantial variation in site size means that when the site size is not associated with impact, the other estimators that do not seek unbiased estimation can substantially outperform. (Compare the Finite Site and Superpopulation Site average columns of Table 2). In these cases, the person-weighted fully uses the data for no cost.

That being said, if there is such an association, using a person-weighted estimator for a site estimand can increase overall error due to the introduced bias. Note the ranges for DB-Persons and FE extend beyond 100, showing that there are cases where the precision gains are more than offset by bias.

The FIRC and RIRC models are generally superior to DB-Sites, but can also be outperformed if there is substantial cross site variation coupled with a size by impact correlation. The overall gains are partially due to the adaptive nature of these estimators: when there is no measured cross-site impact variation they tend towards the more stable precision-weighted estimators, but when there is cross-site impact variation they adapt to target the

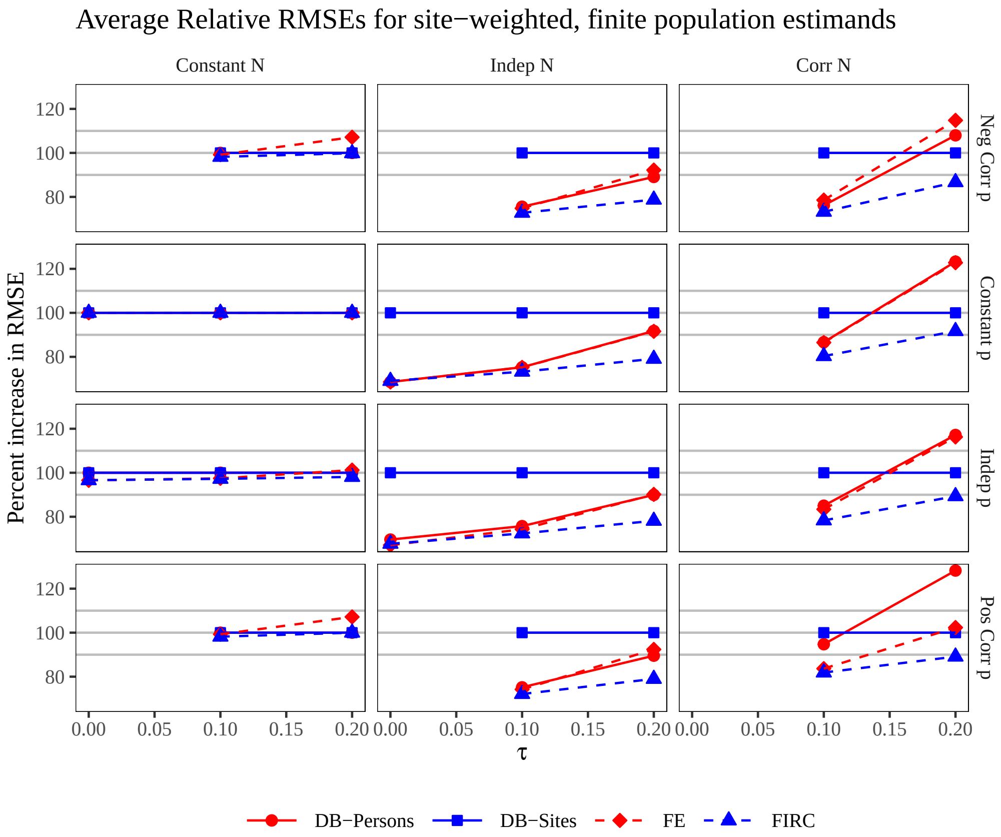

Figure 4. Relative percent increase of RMSE to finite-population site-weighted design-based estimator.

site-average effect. When there is substantial cross-site impact variation correlated with site size, the FIRC and RIRC models still tend towards the person-weighted, giving enough bias to offset their lower SEs. This only occurs in the more extreme scenarios (less than 2% of them, which is why the site relative performance ranges on Table 2 only go up to 100%).

The site-weighted superpopulation stories are roughly the same as finite sample.

2.3 How are the methods differently biased?

We briefly touch on estimator bias. We measure average bias as the mean of the errors across all iterations for a scenario. Bias is the same for finite-sample and superpopulation contexts, so we only list two sets of bias results. See Table 3.

Figure 6 verifies that as cross-site impact variation increases the potential for bias also increases. The biases for FIRC when the correlation with proportion treated is positive or negative shows that the FIRC model is also precision weighted, just like fixed effects. We also see some cases where the biases compound, as illustrated by the larger biases in the double-correlated contexts. The biases can also cancel: see the smaller biases of using the fixed effect estimator for estimating site average impacts when the size correlation is positive and the proportion treated correlation is also positive.

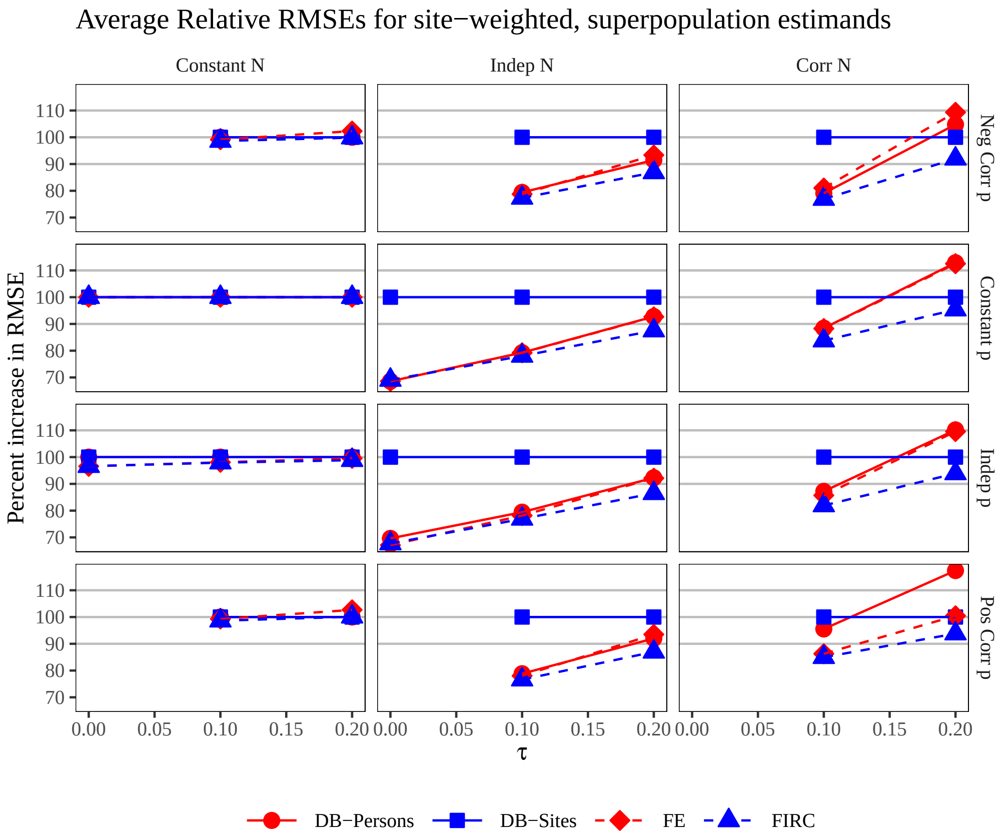

Figure 5. Relative percent increase of RMSE to superpopulation site-weighted design-based estimator. The FIRC method always has lower average RMSE than DB-Sites (there are a small number of high-bias, high-precision scenarios where it is worse). The bias of person-weighting can cause the person-weighted estimators to be worse than DB-Sites, but only in the scenarios with the most extreme cross-site impact variation or scenarios with site impacts correlated with proportion treated (for FE).

	Absolute Bias				Bias as Fraction of RMSE			
	Person		Site		Person		Site	
method	Mean	Range	Mean	Range	Mean	Range	Mean	Range
DB-Persons	0		0.01	(-0.00-0.04)	0	(0-0)	7	(0-36)
DB-Sites	-0.01	(-0.04-0.00)	0		5	(0-30)	0	(0-0)
FE	0.00	(-0.01-0.01)	0.01	(-0.01-0.03)	1	(0-6)	7	(0-34)
FIRC	0.00	(-0.01-0.00)	0.01	(-0.00-0.02)	2	(0-12)	4	(0-22)

Table 3. First sets of columns show bias of estimators across scenarios. Second set show proportion (as a percent) of overall superpopulation RMSE that is due to bias, across scenarios. We see that in some scenarios the (squared) bias is a significant fraction of the overall mean squared error. Intervals are the middle 90% of values.

We also examine what percent bias is of the overall superpopulation RMSE, which also informs how confidence interval coverage would go. We do this by calculating

$$\text{proportion bias} = 100\% \times \frac{\text{bias}^2}{RMSE^2}.$$

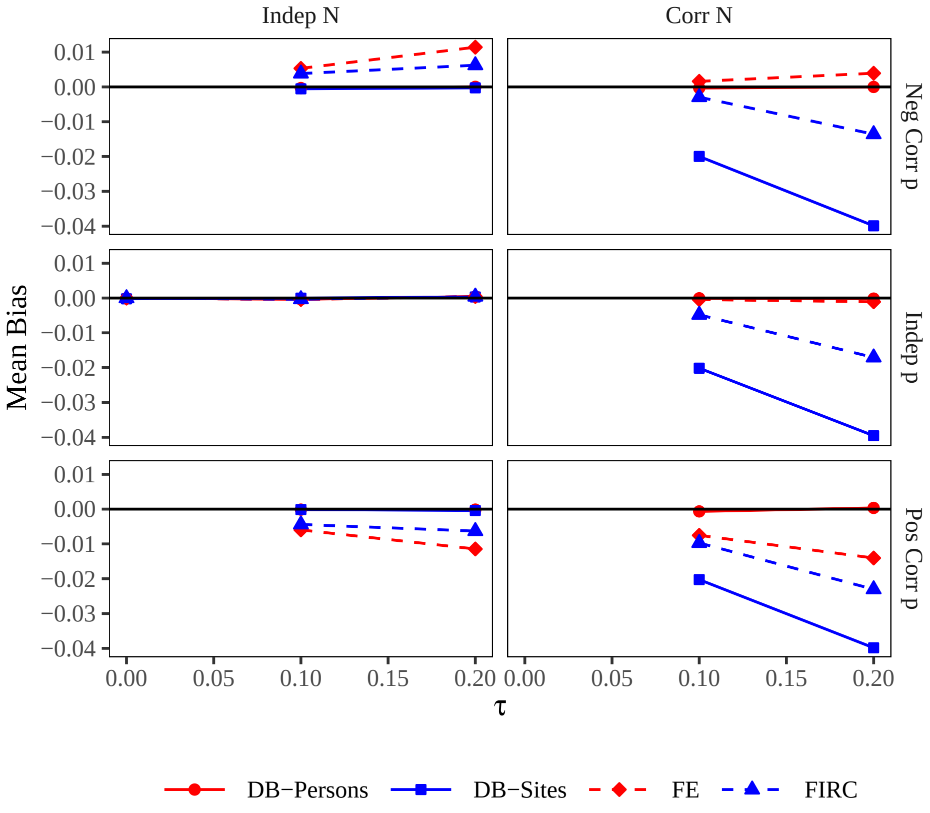

Figure 6. Bias for estimators with respect to person-weighted estimand. Site-weighted biases have corresponding patterns and are therefore not shown. Scenarios with no variation in site size or proportion treated not shown as the biases are similar to independence.

Across all scenarios we have relative superpopulation biases as shown on the right hand columns of Table 3. When using a person-weight targeting estimator for a site-weighted estimand, or vice-versa, bias can be 30% of the overall variation, or more. The bias of FIRC, when taken as an estimator for the site-average impact (as it should be) can reach 22% or more, in the worst case; the shrinkage towards person-weighted can matter substantially. FE, by contrast, has bias of only 6% in the worst cases for the person-weighted estimand, suggesting the consequence of the bias-precision tradeoff is generally not going to be substantive here.

2.4 RICC and RIRC

RICC and RIRC are both nearly identical to FE and FIRC, respectively, and were thus dropped from the above analysis to simplify the presentation of results. Their relationship to the unbiased estimators is nearly identical to those of their sisters, FE, and FIRC. In this section we discuss the differences a bit more explicitly.

2.4.1 RICC

RICC is basically FE, but with replacing the fixed effect intercept with a random effect intercept. The point estimates correlate very highly across all simulation runs, with less than 5% of all simulation runs differing by more than 0.01 effect size units. The largest differences were less than 0.05 effect size units, and 50% of differences were less than 0.002.

We also calculated the relative RMSE across scenarios. RICC method slightly outperformed the FE estimator, with RMSEs on average of 99.7% of the FE (this slight reduction is, in fact, statistically significant, indicating a consistent but tiny improvement). The largest improvements were around a 2% reduction.

2.4.2 RIRC

RIRC and FIRC differ slightly more than RICC vs FE, but the differences are still far from substantial. These estimators differ by more than 0.01 around 15% of the time, and more than 0.02 only around 4% of the time. That being said, for some scenarios these differences reached upwards of 0.15. Overall the RMSE performances are quite similar, with RIRC being potentially slightly worse than FIRC, with RMSEs at most 0.4% higher on average.

2.5 Discussion

Overall, we have examined scenarios with large amounts of variation in site size and the proportion of units treated. This variation does mimic that found in many of our empirical examples, however. This variation provides opportunities for the precision-weighted (e.g., FE) and shrunk (e.g., FIRC) estimators to realize their gains. On the other hand, if this variation is correlated with impact, as it is in many of the scenarios considered, the bias can offset those gains.

Overall, assuming the estimator and estimand are well selected, we find it is difficult to generate scenarios where gains from fixed effects are larger than 10%, but easy to generate ones where the cost is considerable, even beyond the cost due to a mismatch between a person-weighted estimator being used to target a site-weighted estimand, or vice-versa.

On the other hand, the site-weighted design-based estimator is very susceptible to small sites, making it quite unstable when site size substantially varies. The gains of unbiasedness, when site size is correlated with impact, generally appear to be more than offset by this cost, thus suggesting FIRC is a superior choice in general.

3 Performance of the estimated standard errors

In Section 2 we examined the actual performance of the estimates of the Average Treatment Effects. In this section we examine how well the uncertainty (the true standard error) is itself estimated. We now re-introduce the full range of estimators, as even for a given point estimate there can be many ways to calculate a standard error (e.g., cluster robust, heteroskedastic robust, or classic SE estimates for the fixed effect model).

3.1 Are the standard errors calibrated?

We first investigate whether the standard errors of a method equal, on average, the true (appropriately chosen) standard error. We find that yes, all the methods roughly have standard errors close to the appropriately selected true standard errors.

In Figure 7 we, for each method, compare the average estimated standard errors to the actual superpopulation standard error and the actual finite sample standard error. These boxplots show the distribution of the ratios across all the simulation scenarios. We see that the inflation factor generally hovers around 1 (corresponding to the average estimated standard error being the same as the true standard error) when compared to the appropriate context (finite or superpopulation).

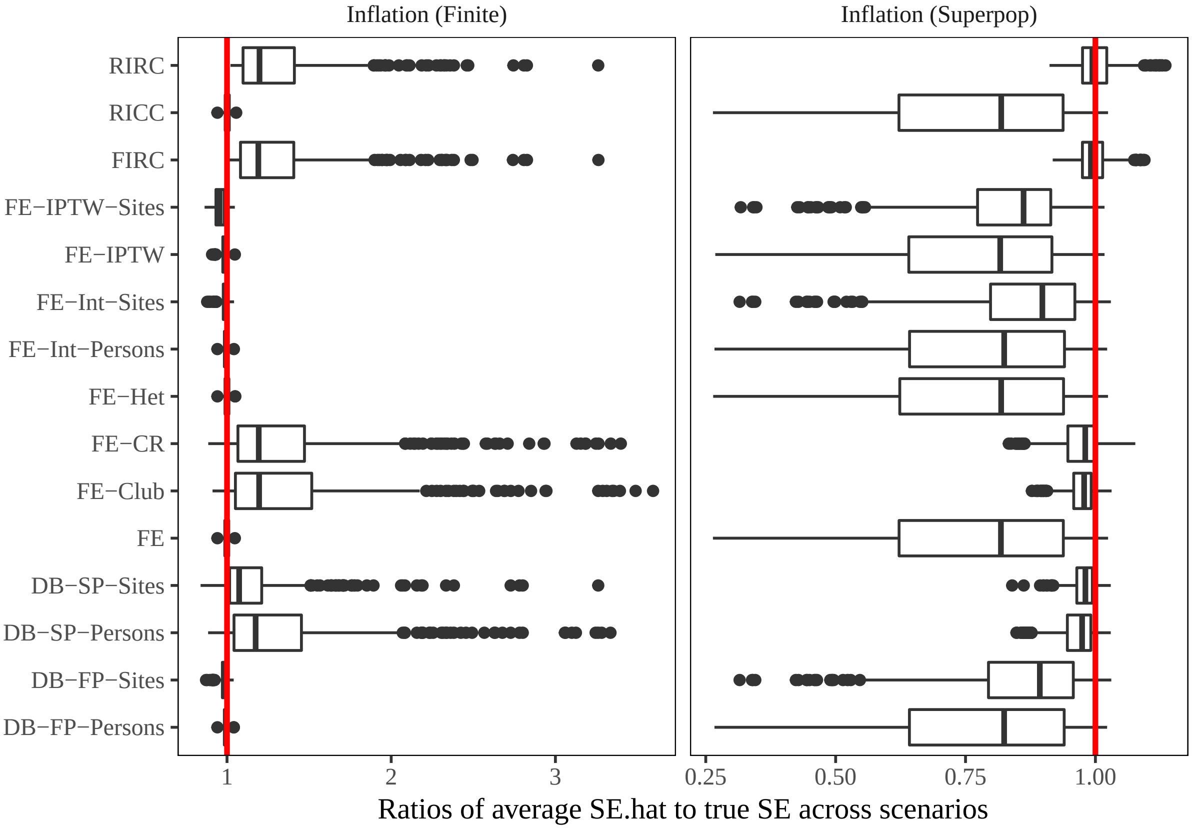

Figure 7. On left, comparing the average estimated standard error to the true finite-sample standard error, averaged across datasets within each scenario. At right, comparing to the true superpopulation standard error. Note that standard errors are independent of person-weighted vs. site weighted choices.

For finite-sample inference, while most standard errors are very well calibrated, some are not as well calibrated as others. In particular, the standard errors for FE-IPTW-Sites tend to be too small, as do those for FE-Int-Sites and DB-FP-Sites. We also see that interpreting a superpopulation method's standard error as the standard error for a finite sample estimand is quite conservative.

For superpopulation targeting estimators, Figure 7 shows trends towards a bit too small for the cluster-robust and design based superpopulation methods, and generally correct calibration overall for FIRC and RIRC. To unpack when this occurs, see Figure 8, where we show the ratio of average estimated standard error to true standard error for the superpopulation targeting estimators as a function of cross-site impact variation. First, we see that the standard errors generally tend to be a bit too small across scenarios (sometimes by more than 10% or more), with the exception of FIRC and RIRC when there is no cross-site impact variation. Also note how FIRC and RIRC, which are actually conservative when there is no cross-site variation, give estimates that are increasingly too small in the face of increasing cross site variation. This is likely a consequence of the pooling; with little variation, FIRC and RIRC's actual SEs are smaller than we would expect (as we saw above) due to tending towards the person-weighted estimates that do not take cross-site variation into account, making the ratios greater than 1. With greater variation this appears to dissipate.

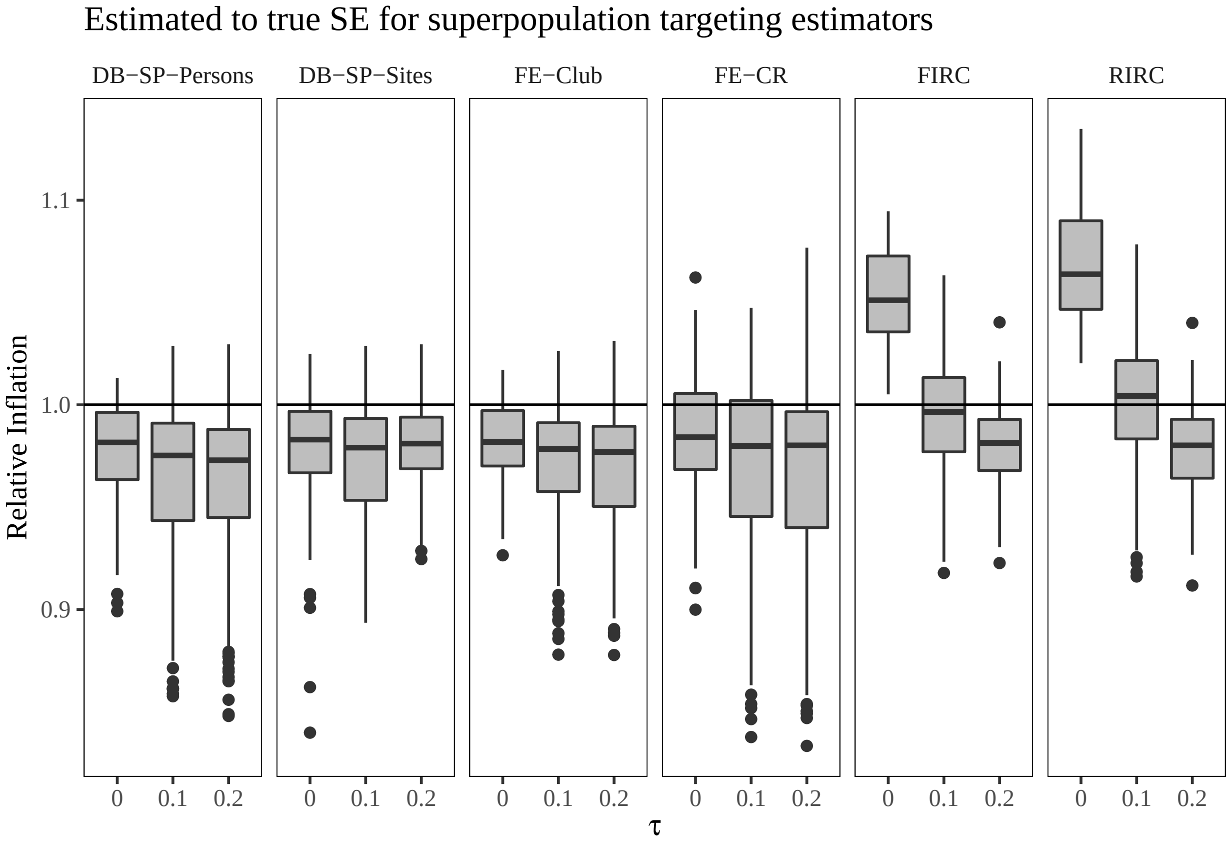

Figure 8. Ratio of estimated SE to true superpopulation population SE for super-population targeting estimators for different degrees of cross site variation. 1.0 corresponds to correct ratio.

Using SEs for the wrong context. We finally briefly discuss how using a finite-population SE as a superpopulation, or vice-versa, can lead to trouble. First, the finite sample standard error methods systematically underreport (usually by around 10%, but often 20% or more, in the scenarios considered) the degree of uncertainty for the superpopulation estimand. The degree of inflation (or deflation) depends on how variable the finite samples are within a given superpopulation context, which is tied to the amount of cross site variation present. In particular, the right sides of the boxplots on the left of Figure 7 (and the left sides of the boxplots on the right) come from those scenarios with maximum cross-site impact variation.

Symmetrically, for super-population inference, using a finite sample standard error (e.g., from the fixed effect model) as a superpopulation one is going to be overly optimistic, and lead to poor coverage. This optimism increases as cross site variation increases. If there is no cross site variation, there is no bias in the SE estimates as all samples have the same target estimand. In the $\tau = 0.2$ scenarios, the SEs can easily be 40% or more too small.

3.2 Are the estimated standard errors differently precise?

For each method, we look at the standard deviation of the standard error estimates for each method to assess how stable these standard error estimates are. These are the SEs of the $\widehat{SE}$s. We can do this nested or overall, to look at both finite-sample and superpopulation performances.

We first calculate the ratio $stdev(\widehat{SE})/SE$, the standard deviation of the estimated standard errors to the true standard error. We plot these ratios across simulation scenarios for each estimator on Figure 9. (For each estimator we only plot for the appropriate population context.) To summarize Figure 9, we also aggregate these ratios across scenarios to create Table 4.

We see that the stability of the standard error estimates is relatively much more unstable for RIRC, FIRC, FE-CR, and the two superpopulation design based methods. The site targeting estimators are also more unstable than the person-targeting. We leave determining whether this instability is compensated for correctly by degrees of freedom adjustments when generating confidence intervals or doing hypothesis testing for future work, but certainly it points to lower power, if not invalidity.

The classic fixed effect methods (including the effectively similar method of RICC) have little to no variability in their standard error estimates. This is likely due to the standard error estimates being primarily dictated by a design matrix that is fixed across randomizations up to permutations, nested by site, of the rows (i.e., the only moving part of the design matrix is the column of treatment assignment dummies). The rest of the variability is a function of the pooled standard deviation estimate, which will generally be quite stable as it assumes cross-site homogeneity. By contrast, the Huber-White robust estimator is more variable as it does not impose a strong homosedasticity assumption and thus cannot take full advantage of the structure of the classic FE approaches. In our simulations we actually *have* homoskedasticity, so this gain is appropriate. We do not explore how wrong the SEs could be in the presence of residual heteroskedasticity.

3.3 How often do the superpopulation estimators give lower SEs than finite population?

Superpopulation estimators have more variable uncertainty estimates, which allows them to quite frequently have smaller estimated SEs than a corresponding finite-population estimator in scenarios with small amounts of cross-site impact variation. For each scenario we counted what proportion of the time a superpopulation estimator's estimated SE was smaller than the DB-Finite-Persons' estimated SE. And to see if these differences are substantial, we also

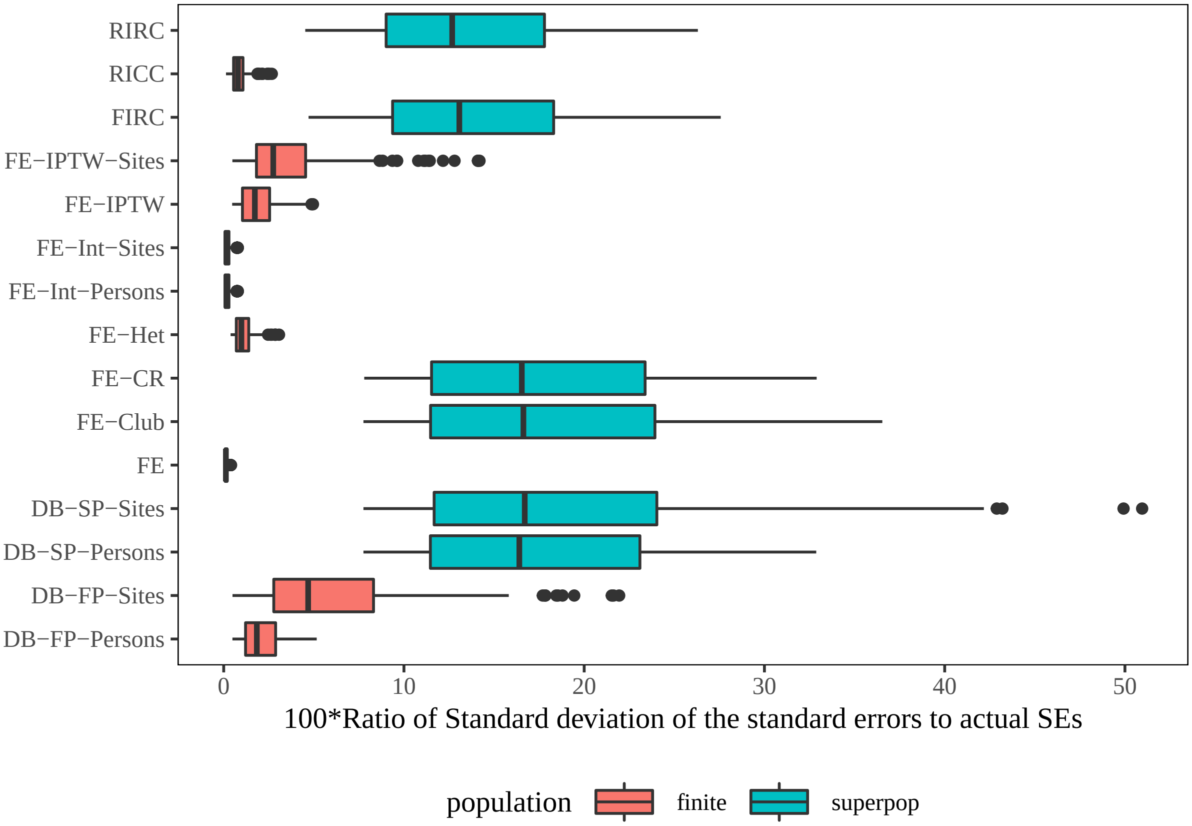

Figure 9. Ratio of standard deviation of $\widehat{SE}$ estimates to true SE across methods.

plot the proportion of time the nominal superpopulation estimated SE was 10% or more smaller. See figure below. We see that, even with cross site variation amounts of around 0.10, we frequently see superpopulation standard errors being smaller more than 25% of the time. For very precisely estimated sites (more sites, or larger sites), this is less of a concern.

The cluster-robust estimator, in particular, is quite unstable and can return very low standard errors. The FIRC model, however, regularly returns lower standard errors, but very rarely undershoots the design based estimator by more than 10%.

3.4 Discussion

Overall, we have seen that the standard error estimates for the finite sample estimators are generally stable and well calibrated. We do not uncover any serious issues, with the possible exception of the inverse probability of treated site weighted estimator.

The superpopulation standard errors, on the other hand, are more difficult to estimate and there are some biases in estimation. In fact, we have seen that these estimators can be so unstable that we can actually end up with estimates that are smaller than the corresponding finite sample estimates in some cases, even though the true finite sample uncertainty is always less.

method	weight	population	Mean	Range
DB-FP-Persons	person	finite	2.1	(0.52-4.08)
FE	person	finite	0.1	(0.02-0.34)
FE-Het	person	finite	1.1	(0.51-1.98)
FE-Int-Persons	person	finite	0.2	(0.03-0.73)
FE-IPTW	person	finite	1.8	(0.52-3.67)
RICC	person	finite	0.9	(0.41-1.51)
DB-FP-Sites	site	finite	5.8	(0.74-12.88)
FE-Int-Sites	site	finite	0.2	(0.03-0.73)
FE-IPTW-Sites	site	finite	3.5	(0.75-8.29)
DB-SP-Persons	person	superpop	18.0	(8.32-30.51)
FE-Club	person	superpop	18.4	(8.23-33.10)
FE-CR	person	superpop	17.9	(8.36-30.31)
DB-SP-Sites	site	superpop	19.2	(8.32-35.76)
FIRC	site	superpop	14.4	(6.60-25.33)
RIRC	site	superpop	14.0	(6.29-24.48)

Table 4. Average and range across scenarios of the degree of instability in estimating the standard error as a percent of true standard error

4 Coverage

We calculated nominal 95% confidence intervals directly using the standard errors and point estimates for each method using the ±1.96 rule (i.e., assuming normality rather than adjusting for any degrees of freedom). We then compared the estimand targeted by the method to this interval, scoring a success for each time the true parameter lay inside the interval. We then calculate the percent of successes for each method for each scenario. This means each coverage rate is specific to a particular population and weighting.

Results are on Figure 11. Coverage can be low, but even in the worst case we see rates of at least 85%. The low tails for the fixed effect methods are due to bias in estimation, not inappropriately calculated standard errors. We generally see the superpopulation methods have low coverage, especially for small number of sites. This is partially due to not correcting for degrees of freedom with a t-distribution rather than normal, although this should not be a concern for 40 sites or more. Some of the coverage should be due to the bias due to fixed-effects in the case of the FE-CR and FE-Club estimators, which makes the similarity of their coverage with DB-SP-Persons quite interesting. For superpopulation, site-average we see the bias of RIRC and FIRC hurting coverage, while the design-based does not have that particular concern.

As a final check, we look at coverage for those scenarios where there should be no bias due to a correlation of site impact with either site or proportion treated. See Figure 12. We see generally good coverage although superpopulation estimators still need further adjustment via more appropriate degrees of freedom corrections.

As a final comment, we note that one can obtain good coverage even with very poorly estimated standard errors. If the estimator returns very large values some of the time, then

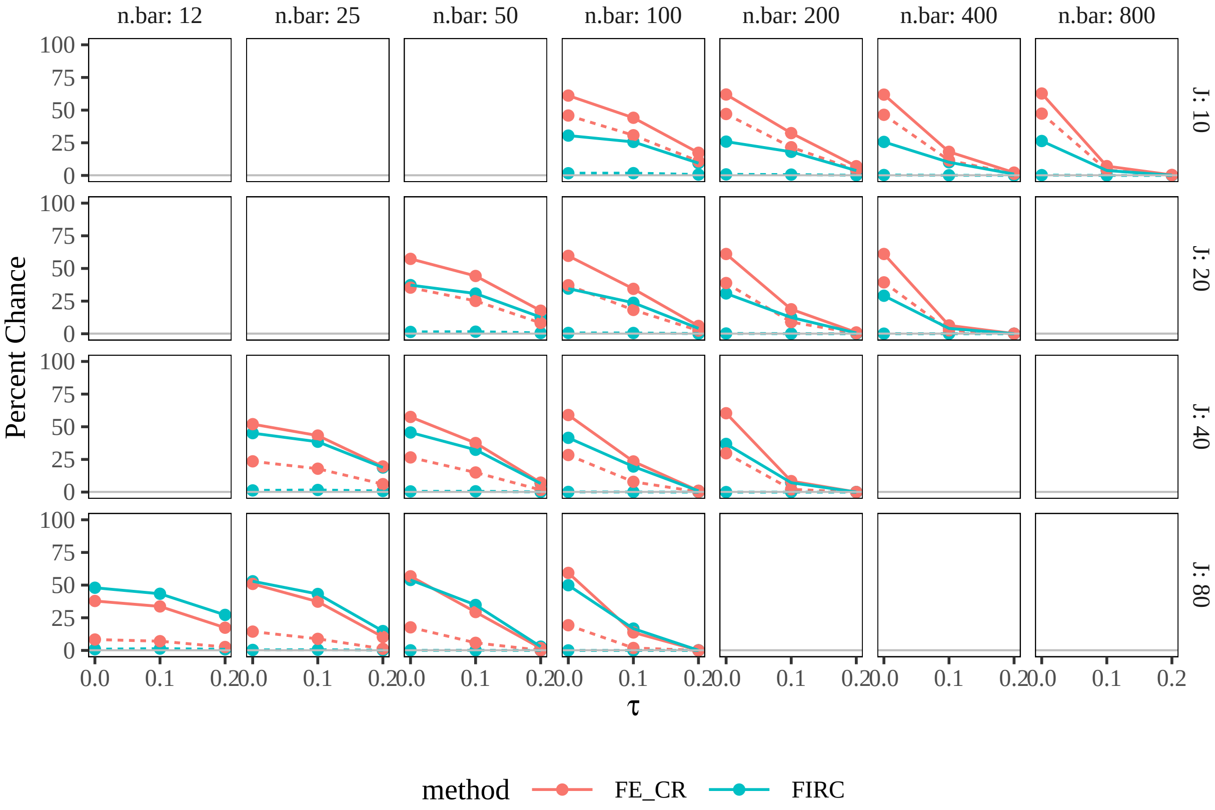

Figure 10. Proportion of time superpopulation estimator has smaller $\widehat{SE}$ than finite-sample design-based estimator. The dotted lines show proportion of time it is 10% smaller or more.

no matter how poor the point estimate, the confidence interval will cover. This can offset the standard error being too small for other times. (The t-distribution is used to account for this in some contexts.) This arguably is what is occurring for some of our superpopulation estimators.

Future work should more throughly investigate the above with appropriate degrees of freedom adjustments. These would lead to lower power and wider intervals. Overall, this further suggests that not only are superpopulation targeting estimators more unstable in truth, in the face of cross site impact variation, but that measuring and testing in this context has an additional layer of difficulty on top of that.

5 Comparing FIRC to DB-Super-Sites

In our empirical analysis we identified the most amount of variation in estimated standard errors within the domain of the site-weighted, superpopulation estimand. The two estimators in this domain are FIRC and DB-SP-Site. We wanted to investigate to what degree these differences are due to different *actual* standard errors, or different instability in *estimating* standard errors. In this section we therefore look at all the above results with an eye to comparing how FIRC, which is adaptive in that it tends towards precision-weighted to stabilize its estimates of the superpopulation site average impact, performs relative to the

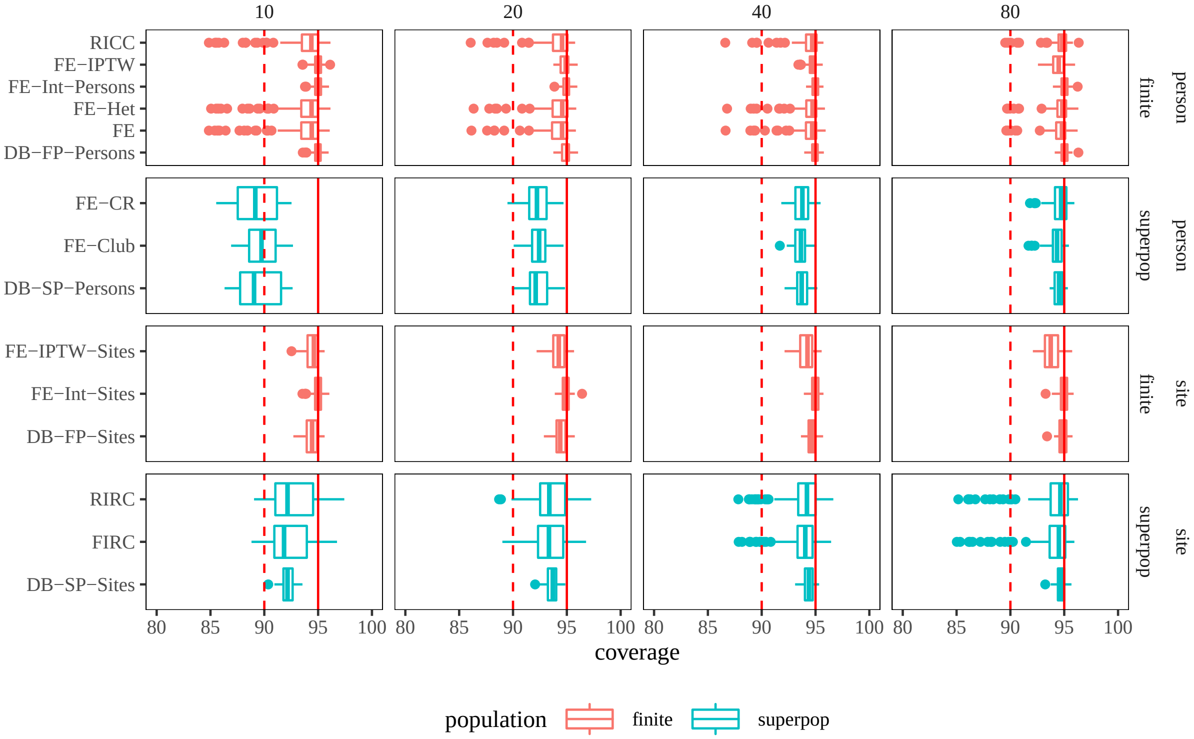

Figure 11. Percent of nominal confidence intervals that capture the true method-appropriate estimand across scenarios. Plots grouped by target estimand. Plots left to right are for different number of sites. Monte-carlo standard errors for individual coverage rates are 1 percentage point or smaller.

unbiased DB-SP-Sites. Overall we find that FIRC is more stable, although DB-SP-Sites can have better coverage when there is actual bias.

First, for many scenarios the true SEs of $\hat{\beta}_{FIRC}$ and $\hat{\beta}_{DB-SP-Site}$ were nearly the same, but the design-based estimator became substantially more variable than FIRC in the face of variation in site size. The average increase was 17%, averaged across all scenarios including those with no variation in site size or proportion treated. Several scenarios (generally when there was little to no cross site variation) had increases of 50% or more.

FIRC is making a bias-variance trade-off, and thus we might imagine that the RMSE for FIRC will be higher than DB-SP-Site for those scenarios where site size is correlated with site impact. In general, we found that across of all simulation scenarios we consider, the RMSE of FIRC was higher than DB-SP-Site in only 2% of them.[4] We find that DB-SP-Site can modestly outperform FIRC in those cases with many sites ($J = 80$), maximal variation $\tau = 0.20$, and site size correlated with impact. Overall, even in those contexts where FIRC is facing bias, it generally has a superior RMSE as the savings of a more stable SE significantly offset the bias cost. This does mean that coverage of FIRC can be worse than DB; see coverage, below for further discussion.

[4] To further account for Monte Carlo estimation error, we fit a linear regression to determine when FIRC would be inferior. Our estimated percent remained just below 2%.

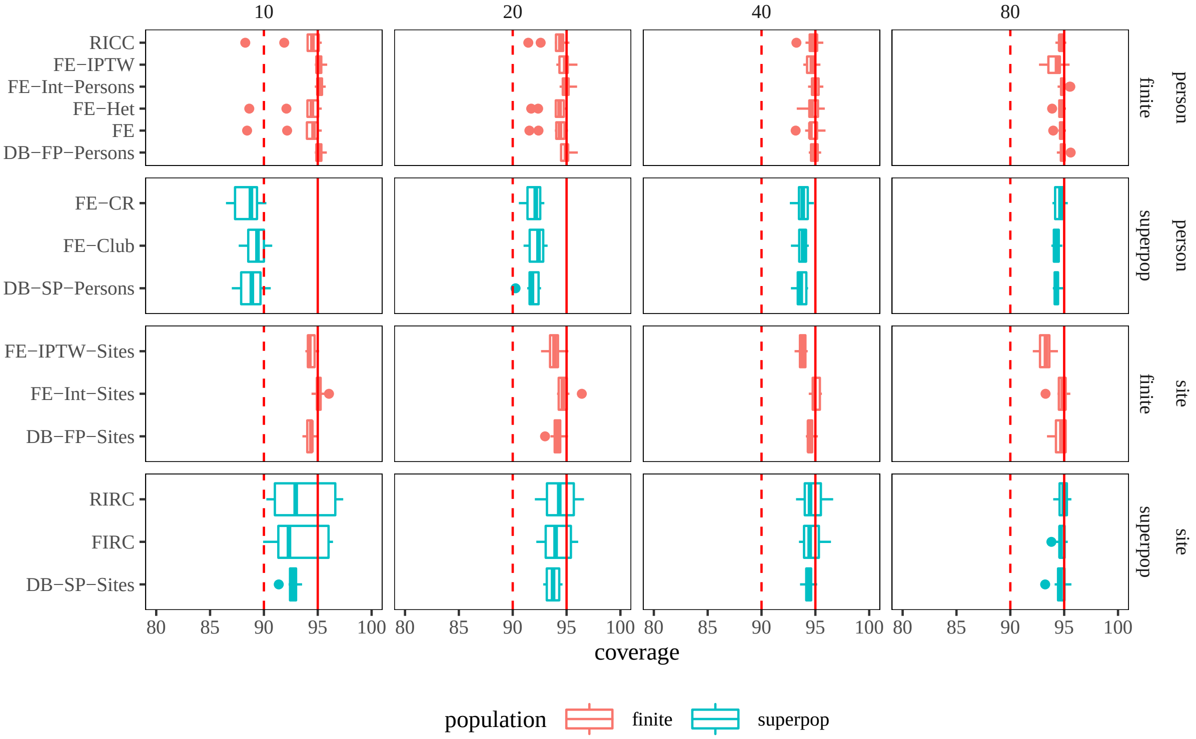

Figure 12. Analog of Figure 11, but only for scenarios without bias due to correlation of impact and site size or proportion treated. Monte-carlo standard errors for individual coverage rates are 1 percentage point or smaller.

Because the FIRC true standard errors are generally smaller than DB-SP-Sites, comparing the uncertainty in the estimated standard errors directly is potentially confusing. We instead compare the relative proportions of between the two estimators, and find that the estimated SEs are, on average, 34% more unstable for DB-SP-Site than FIRC. When it comes to standard error estimation, FIRC is almost uniformly preferable to DB-SP-Sites.

We visualize both these measures, RMSE and relative stability in estimating the standard error, in Figure 13. We plot each scenario in terms of the relative performance with regard to RMSE vs. estimating the SE, and find that FIRC generally outperforms in both measures. We also see that when one is much superior, the other generally is as well. We also see that estimation of SEs is still problematic for DB-SP-Sites even when sites are the same size.

We finally compare ecoverage of the two estimators. See Figure 14. We see that when there is bias, FIRC's superior RMSE and estimated SEs do not save it from poor coverage. Here, because the estimated SEs are estimating variablility, not bias, FIRC can miss its mark more than DB-SP-Site; see how the points tend towards 95% for DB-SP-Site but not FIRC in the bottom right case of high cross site variation correlated with site-level ATE.

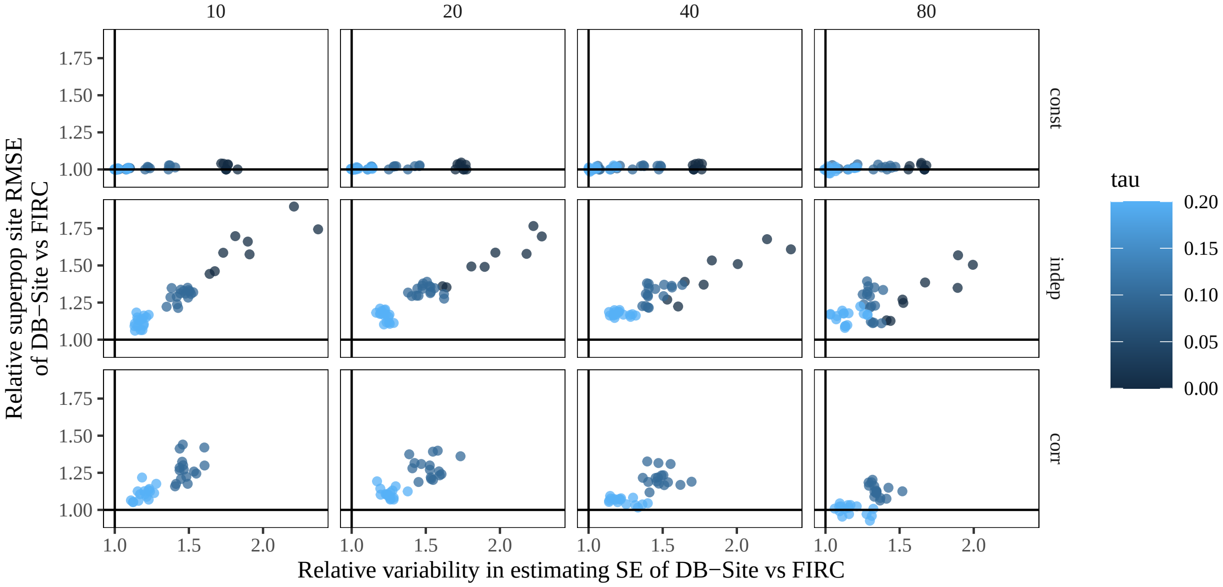

Figure 13. Relative performance in estimating SEs vs relative performance in RMSE for FIRC vs. DB-Site (both targeting the superpopulation site estimand). We see FIRC generally outperforms on both metrics except with many sites and large degrees of variation correlated with site size.

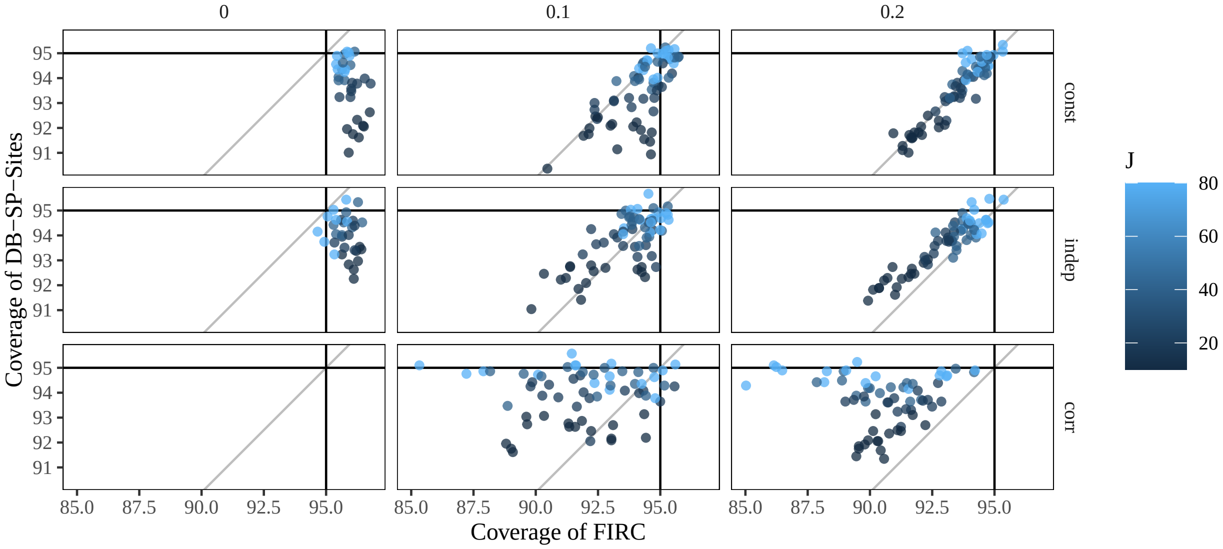

Figure 14. Coverage of FIRC and DB-SP-Site. The tiles show increased treatment variation from left to right, and three types of site variability (constant, independent of impact, and correlated with impact) from top to bottom. Points along the 95 lines are ideal. Above means overcoverage, and below means undercoverage.