Characterizing Cross-Site Variation in Local Average Treatment Effects in Multisite RDD contexts with an Application to Massachusetts High School Exit Exam

Sophie Litschwartz
Harvard Graduate School of Education

Luke Miratrix
Harvard Graduate School of Education

March 16, 2021

Abstract

In multisite experiments, we can quantify treatment effect variation with the cross-site treatment effect variance. However, there is no standard method for estimating cross-site treatment effect variance in multisite regression discontinuity designs (RDD). In this research, we rectify this gap in the literature by developing and validating a method based on random effects meta-analysis. We also evaluate two fixed intercepts/random coefficients (FIRC) models based on prior RDD studies and use simulation to demonstrate their unsuitability. We then apply our model to a high school exit exam policy in Massachusetts that required students who passed the high school exit exam but were still determined to be nonproficient to complete an "Education Proficiency Plan" (EPP). We find the EPP policy had a positive local average treatment effect on whether students completed a math course their senior year, but that the treatment effect variance was large enough in three cohorts for the treatment effect to have been negative in more than a third of high schools.

Introduction

In Massachusetts, students who score as "Need Improvement" but not "Proficient" on the ELA and math high school exit exam, which students first take at the end of 10th grade, are required to complete an "Education Proficiency Plan" (EPP) before they can graduate. This policy is a classic set up for a regression discontinuity analysis. Students are assigned to treatment (i.e., being required to complete an EPP) based on whether their value on a running variable (i.e., 10th grade high school exit exam score) falls above or below a specific cut point (i.e., scoring above or below the minimum proficiency score). For students who tend to score near the cut-point measurement error in the running variable makes assignment into the EPP near random (Imbens & Lemieux, 2008; Lee & Lemieux, 2010), and the causal effect of EPP can be estimated by comparing the outcomes of students just below the cut point to students the outcomes just above the cut point. In our case, we find that being required to complete an EPP had an overall local average treatment effect of increasing the probability a student completes a math course in their senior year of about 3 percentage.

However, estimating just the local average treatment effect does not provide a full picture of the EPP policy's effects. The EPP policy could have the same effect in all high schools, or the effect could vary considerably across schools, with the effect being large in some and small or even negative in others. Quantifying treatment effect variation is therefore important for understanding the full range of expected policy impacts, where and with which populations a policy is most effective, and how generalizable policy effects are outside the study sample (Angrist, Pathak, & Walters, 2013; Raudenbush & Bloom, 2015; Tipton, 2014; Weiss, Bloom, & Brock, 2014).

One measure of treatment effect variation in multi-site studies is the cross-site treatment effect variance. There are standard methods for estimating cross-site treatment effect variance in multisite randomized experiments, but there are no such methods designed to be used in RDDs. This is despite the fact that RDDs have all the same sources of treatment effect variation as randomized experiments. In fact, the conditions under which people most often use RDDs are precisely the conditions under which we see the most cross-site variance within random controlled

trials (RCT): interventions which are only loosely specified (Weiss et al., 2017). RDD, a quasi-experimental method, is most often used opportunistically to study policies where interventions were naturally assigned using a cut-score on a running variable. Such natural experiments generally have interventions that are less tightly controlled and specified than RCTs that are pre-planned and implemented by researchers.

In the first part of this paper, we develop and evaluate a method for estimating cross-site treatment effect variation based on random effects meta-analysis. We treat our multisite study as a form of "planned meta-analysis" (Bloom, Raudenbush, Weiss, & Porter, 2017). In each site we run a local linear regression model using only data from that site to estimate a site-level treatment effect. These site-level treatment effects are then combined to get an average treatment effect and a cross-site treatment effect variance using tools from random effects meta-analysis (Higgins, Thompson, & Spiegelhalter, 2009).

We also evaluate two other methods based on analyses done in the only other studies we are aware of to estimate the cross-site treatment effect variance in a multisite RDD: a study by McEachin, Domina, and Penner (2020) on the effect of early algebra in California middle schools and a study by Raudenbush, Reardon, and Nomi (2012) on statistical methods for multisite trials that includes an evaluation of double dose algebra in Chicago schools. Both of these studies use a fixed intercepts random coefficient (FIRC) model, where outcomes are modeled with a multi-level model with a fixed unpooled intercept for each site and a site-level random effect coefficient on treatment (Bloom et al., 2017). However, each of the two studies makes different choices in how to model the running variables in the equation.

Coefficients in a regression framework can be modeled one of three ways: pooled, unpooled, and partially pooled. A pooled estimate is a typical ordinary least squared coefficient, where all data across sites is combined to estimate one coefficient value for all sites. An unpooled estimate is where the coefficient is estimated only using data from each individual site and there is a different estimate for each site. A partially pooled coefficient is modeled as a random effect, where the coefficient is assumed to be normally distributed across the sites.

McEachin et al. (2020) model the running variable with pooled coefficients on the running variable and the treatment running variable interaction. The pooled coefficients for the running variable terms simplifies the model but also can leave the model misspecified if there is cross-site variation in these coefficients. Raudenbush et al. (2012) have a more complicated model with a random effect on running variable term. Their model also has a quadratic running variable term instead of a linear interaction between the treatment status and the running variable. However, using a linear treatment running variable interaction is more in line with contemporary RDD standards (Gelman & Imbens, 2019), and so we adapt the Raudenbush et al. (2012) model by replacing the quadratic running variable term with a linear treatment running variable interaction. We model the coefficient on the linear treatment running variable interaction as a partially pooled random effect.

Neither McEachin et al. (2020) or Raudenbush et al. (2012) provide confidence intervals for their cross-site treatment effect variance estimates. We evaluate three potential methods for estimating confidence intervals in an RDD FIRC model: Wald standard errors, Q-statistics inversion, and profiled confidence intervals. Using simulation, we demonstrate both that Wald standard errors and Q-statistics inversion are problematic when used with the RDD FIRC model, leaving profile confidence intervals as the most reasonable confidence intervals for cross-site treatment effect variance in the FIRC models.

That being said, our evaluation of the different potential models for estimating cross-site treatment effect variance in multi-site RDDs shows that FIRC models do not work well. The RDD FIRC models frequently suffer from colinearity and have a "singular" fit, making the model estimates unusable. The simplified FIRC model with pooled running variable coefficients produces a singular fit less often, but has a large upward bias when there is cross-site variance in the running variable coefficients. In contrast, the random effects meta-analysis model produces reasonable estimates across a variety of conditions with the point estimate in the 95% confidence interval approximately 95% of the time.

In the second part of the paper, we apply these methods to the EPP policy example. The

EPP policy grants individual high schools across Massachusetts considerable latitude in implementing EPPs for their students. High schools can require students to demonstrate proficiency by taking a special proficiency exam, passing courses in the relevant area(s) in their junior and senior year, or a combination of the two. The high school certifies final proficiency, and the state does not require high schools to make students take the high school exit exam again.

Individual high school implementation decisions are particularly relevant for the math EPP. Massachusetts does not impose ELA or math high school graduation requirements at the state level, but in practice, all Massachusetts high schools require four years of ELA. However, there is variation across Massachusetts high schools in how much math they require, with high schools requiring anywhere between two and four years of math to graduate. With the ELA EPP, the exam has no binding impact on a student's coursework requirements because students are already required to pass four years of ELA to graduate. However, with math, an EPP could increase the number of math courses a student must complete because there is no baseline requirement of four years of math to graduate.

We use our random effects meta-analysis RDD model to understand how high schools implemented the EPP policy in practice. One goal of the policy was to increase the percentage of high school seniors in Massachusetts who completed a math course their senior year. While this worked on average, we find large cross-site treatment effect variance across high schools. We also find that differences in high school graduation requirements are not enough to explain this variance and that the variance across high schools that required four years of math is almost as large as the variance across high schools that did not require four years of math. Therefore we can conclude there were meaningful differences in program implementation across schools beyond differences in course requirements.

Analytical Models

Multisite RDD studies generally estimate causal effects using local linear regression (Hahn, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008; Gelman & Imbens, 2019). This local linear regression model (LLR) for multisite RDDs frequently take the following form for

observation i in site j:

$$Y_{ij} = \alpha_j + \beta_1 T_{ij} + \beta_2(Score_{ij} - Score_c) + \beta_3 T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij}$$
$$\epsilon_{ij} \overset{iid}{\sim} N[0, \sigma_y^2] \tag{1}$$

where Y_{ij} is the outcome of interest, α_j is an unpooled site level intercept, $Score_{ij}$ is the running variable, $Score_c$ is the treatment cut score and T_{ij} is binary treatment indicator determined by whether $Score_{ij}$ is above/below $Score_c$. Generally, the model is estimated only using observations where $|Score_{ij} - Score_c| < h$, where h is the model bandwidth.

In this model, the local average treatment effect is β_1. β_1 is a single fixed treatment effect pooled across sites, which does not allow for the estimation of cross-site treatment effect variance. The confidence interval for β_1 is estimated using clustered robust errors clustered at the site level.

Our proposed method for estimating cross-site treatment effect variance treats the multisite RDD as a random effects meta-analysis of small site-level RDD studies. In each site, a separate regression model is estimated as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j}(Score_{ij} - Score_c) + \beta_{3j} T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij}$$
$$\epsilon_{ij} \overset{iid}{\sim} N[0, \sigma_{yj}^2] \tag{2}$$

Under this model the site specific variance covariance matrices of the vector of coefficients $\widehat{\boldsymbol{\beta_j}}$ would generally be estimated as:

$$VCov(\widehat{\beta_j}) = (X_j' X_j)^{-1} \widehat{\sigma_j}^2$$
$$\widehat{\sigma_j^2} = \frac{1}{n_j - 3 - 1} \sum (Y_{ij} - \widehat{Y_{ij}})^2 \tag{3}$$

where X_j is the data matrix of dependent variables for site j and n_j is the total number of observations in site j.

However, estimating the variance covariance matrix separately for each site can lead to imprecise standard error estimates, especially for small sites. Instead, we increase the precision of the standard error estimates by modeling the coefficient variance covariance matrix using a pooled estimate for residual variance as follows:

$$\begin{aligned} VCov(\widehat{\beta_j}) &= (X_j' X_j)^{-1} \overline{\widehat{\sigma_j}^2} \\ \overline{\widehat{\sigma_j^2}} &= \frac{1}{\sum n_j} \sum (n_j \hat{\sigma_j}^2) \end{aligned} \tag{4}$$

Consistent with the meta-analysis literature (Higgins et al., 2009; Whitehead & Whitehead, 1991; DerSimonian & Laird, 1986), we estimate the overall average treatment effect as a precision weighted average of the site-level treatment effects. This means we estimate the overall local average treatment effect as follows:

$$\widehat{\beta_1} = \frac{\sum \widehat{\beta_{1j}} w_j}{\sum w_j},\ \widehat{SE_{\beta_1}} = \sqrt{\frac{1}{(\sum w_j)}} \tag{5}$$

where the site level weights $w_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}} + \widehat{\sigma^2_{\beta_1}}}$.

The cross-site treatment effect variance $\sigma^2_{\beta_1}$ is calculated using a methods of moments estimator:

$$\begin{aligned} \widehat{\sigma^2_{\beta 1}} &= max(0, \frac{Q - J - 1}{\sum \widehat{SE^{-2}_{\beta_{1j}}} - \frac{\sum \widehat{SE^{-4}_{\beta_{1j}}}}{\sum \widehat{SE^{-2}_{\beta 1j}}}}) \\ Q &= \sum W_j (\widehat{\beta_{1j}} - \frac{\sum W_j \widehat{\beta_{1j}}}{\sum W_j})^2, where\ W_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}}} \end{aligned} \tag{6}$$

where $\widehat{SE_{\beta_{1j}}}$ is the estimated site-level standard error for β_{1j} from Equation 4 and J is the total number of sites.

The confidence interval for the cross-site treatment effect variance is estimated using

Q-statistic inversion. In meta-analysis, the Q-statistic is defined as:

$$Q(\tau^2) = \sum_{j=1}^{J} \frac{(\widehat{\beta_{1j}} - \frac{\sum W_j \widehat{\beta_{1j}}}{\sum W_j})^2}{\widehat{SE^2_{\beta_{1j}}} + \tau^2}, where\ W_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}}} \tag{7}$$

The Q-statistic $Q(\tau^2)$, is similar to the Q in the methods of moment estimator, but $Q(\tau^2)$ includes the treatment effect variance (τ^2) in the denominator (Higgins et al., 2009). This Q-statistic has a chi-squared distribution with J-1 degrees of freedom. Under the test-inversion procedure the Q-statistic is estimated for a plausible range of τ^2 values from 0 to some τ^2_{max}. These Q values are compared to $\chi^2_{J-1}(\frac{\alpha}{2})$ and $\chi^2_{J-1}(1-\frac{\alpha}{2})$, where α is the level of the confidence interval. The α confidence interval for $\sigma^2_{\beta_1}$ is all τ^2 where $Q(\tau^2) \geq \chi^2_{J-1}(\frac{\alpha}{2})$ and $Q(\tau^2) \leq \chi^2_{J-1}(1-\frac{\alpha}{2})$.

In addition to our proposed random effects meta-analysis model, we evaluate two different variations of an RDD FIRC model. In both RDD FIRC models, the treatment effect is modeled with a site-level random effect using a multi-level model. However, in the first FIRC model, both running variable coefficients are pooled across sites, and in the second FIRC model, both running variable coefficients are modeled as site-level random effects along with the treatment impacts.

FIRC One (restricted):

Level One - Observation:

$$Y_{ij} = \alpha_j + \beta_{1j}T_{ij} + \beta_2(Score_{ij} - Score_c) + \beta_3 T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij}$$
$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_y)$$

Level Two - Site:

$$\beta_{1j} = \delta + e_{1j} \tag{8}$$
$$e_{1j} \overset{iid}{\sim} N(0, \sigma^2_{\beta_1})$$

FIRC Two (unrestricted):

Level One - Observation:

$$Y_{ij} = \alpha_j + \beta_{1j}T_{ij} + \beta_{2j}(Score_{ij} - Score_c) + \beta_{3j} T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij}$$
$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_y) \tag{9}$$

Level Two - Site:

$$\beta_{1j} = \delta + e_{1j}$$
$$\beta_{2j} = \gamma_2 + e_{2j}$$
$$\beta_{3j} = \gamma_3 + e_{3j}$$
$$\begin{pmatrix} e_{1j} \\ e_{2j} \\ e_{3j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \sigma^2_{\beta_1} & \sigma_{\beta_1\beta_2} & \sigma_{\beta_1\beta_3} \\ \sigma_{\beta_2\beta_1} & \sigma^2_{\beta_2} & \sigma_{\beta_2\beta_3} \\ \sigma_{\beta_3\beta_1} & \sigma_{\beta_3\beta_2} & \sigma^2_{\beta_3} \end{pmatrix} \right]$$

Both of these models are estimated only using observations within a set bandwidth away from the cut score, and in both cases, δ represents the local average treatment effect.

Neither McEachin et al. (2020) or Raudenbush et al. (2012) estimate a confidence interval for their cross-site treatment effect variance estimate, so we test three possible methods for obtaining confidence intervals: 1) Wald standard errors, 2) Q-statistic inversion, and 3) profiled confidence intervals. Wald standard errors are known to be unreliable for variance parameters in multi-level models and are particularly problematic when variances are close to zero; however, we present them for evaluation purposes. Following (Bloom et al., 2017), the Q-statistic confidence intervals are not estimated using the parameters from the multi-level model but an analogous OLS regression model. Like Q-statistic inversion, the profiled confidence interval also uses test inversion, although now it is the likelihood ratio test being inverted. Different potential cross-site treatment effect variance estimates are plugged into the likelihood functions and the 95% confidence interval is all values where the likelihood ratio test is not significant at the .05 level.

Simulation Specifications

We use simulations to compare how our analytical models perform under different empirical conditions. LLR doesn't allow for the estimation of cross-site treatment effect variance, but we do compare our models to LLR in the case of the average treatment effect.

Under the potential outcomes framework, data for observation i in site j is generated as

follows:

$$
\begin{aligned}
Y_{0ij} &= a_{0j} + b_{0j} Score_{ij} + \epsilon_{ij} \\
Y_{1ij} &= Y_{0j} + a_{1j} + b_{1j} Score_{ij} \\
\epsilon_{ij} &\sim N(0, \sigma^2_\epsilon) \\
Score_{ij} &= \mu_{Score} + r_j + \pi_{ij} \\
\pi_{ij} &\sim N(0, 1 - ICC_{Score})
\end{aligned}
\tag{10}
$$

where Y_{0ij} is the outcome absent treatment, Y_{1ij} is the outcome with treatment, σ^2_ϵ is the residual variance of the outcome, r_j is the site-level effect on the running variable, and ICC_{Score} is the inter-class correlation of the running variable. In this model $Score_c$ is fixed at 0 and the overall running variable distribution is constructed to have a standard deviation of one and a grand mean of μ_{Score}.

Under this data generating process all coefficients are site specific. The residual error variance, σ^2_ϵ, is assumed to be fixed across sites and observations. Each site is also defined as having n_j observations and a site-level mean shift on the running variable of r_j.

The site level parameters are generated as follows:

$$
\begin{aligned}
a_{0j} \sim N(\mu_{a0}, \sigma^2_{a0}), b_{0j} \sim N(\mu_{b0}, \sigma^2_{b0}), a_{1j} \sim N(\mu_{a1}, \sigma^2_{a1}), b_{1j} \sim N(\mu_{b1}, \sigma^2_{b1}) \\
r_j \sim N(0, ICC_{Score}) \\
n_j \sim Pois(\mu_n)
\end{aligned}
\tag{11}
$$

where $\mu_{a0}...\mu_{b1}$ are the coefficient means, $\sigma^2_{a0}...\sigma^2_{b1}$ are the coefficient variances, and μ_n is the average number of observations per site. All the model coefficients are assumed independent from each other.

The data generating process above means that overall our simulation takes as input parameters means and variances for each coefficient $a_0...b_1$, a residual variance value σ^2_ϵ, a value for the interclass correlation of the running variable ICC_{Score}, and the average observations per site μ_n. In addition, the total number of sites J, the bandwidth h, and a running variable grand

mean μ_{Score} is specified for each simulation. In this model, μ_{a1} is the true local average treatment effect and σ^2_{a1} is the true cross-site treatment effect variance.

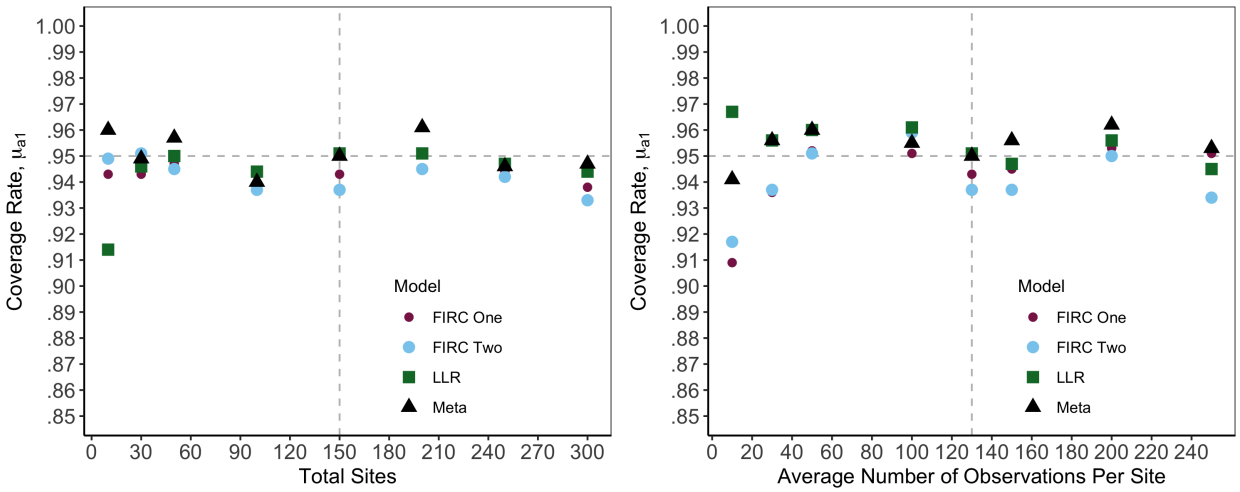
The baseline parameter values are based on the empirical data from Massachusetts. Across all simulations we fix the average parameter values as: $\mu_{a0} = .7$, $\mu_{b0} = .05$, $\mu_{a1} = .07$, $\mu_{b1} = .025$. We also fix the control mean standard deviation (σ_{a0}) to .3, the treatment effect standard deviation (σ_{a1}) to .07, the residual error (σ^2_{ϵ}) to .4, the bandwidth (h) to 1, and the running variable grand mean (μ_{Score}) to 1. Across simulations we vary the average observations per school (μ_n) from 10 to 250 with a baseline value of 130, the total schools (J) from 10 to 300 with a baseline value of 150, the standard deviation of the control running variable coefficient (σ_{b0}) from 0 to .3 with a baseline value of .05, the standard deviation of the treatment running variable coefficient (σ_{b1}) from 0 to .3 with a baseline value of .025, and the running variable ICC (ICC_{Score}) from 0 to .9 with a baseline value of .2. Therefore we run a total of 46 simulations with each simulation parameter besides the parameter being varied fixed at its baseline value.

Simulation Results

Estimate of the Treatment Effect Mean

The LLR model, the random effects meta-analysis RDD, and both the RDD FIRC models produce reasonable estimates for the local average treatment effect. Using the benchmark parameter values, we see that all three models have coverage rates (i.e. the proportion of simulations where the model estimate is in the confidence interval) close to 95% across site sizes and the total number of sites (Figure 1). The only two exceptions are for small sample sizes. When there are only ten sites, the LLR model has a coverage rate of 91%, which is worse than the other two models, and when there is only an average of ten observations per site, the two FIRC models have coverage rates of approximately 91%. For all four models, the extent to which the coverage is below 95% is driven by small underestimates of the standard errors and not bias in the estimate. Overall, these results provide reassuring evidence that the model-misspecification in the LLR model typically used by researchers does not interfere with the estimate of the local

average treatment effect.



(a) (b)

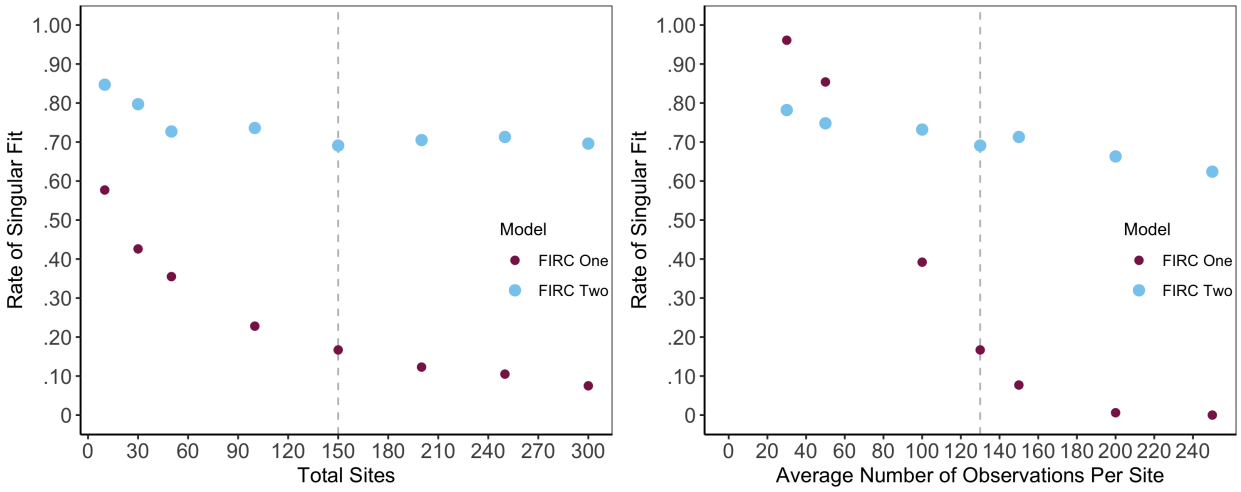
Figure 1: The coverage rate of the local average treatment effect estimates across the local linear regression (LLR), the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity models. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right).

Both FIRC Models Run Into Estimation Problems

Both the FIRC models used in prior research run into a range of problems in our simulations. In general, multi-level models are data intensive and will not run properly without enough clusters or if the clusters are not large enough. Sample size requirements are even larger in FIRC models, which soak up a significant amount of the model variance with the site-level fixed intercepts. The RDD model only uses data near the cutoff, which only further increases the required amount of data. Therefore the RDD FIRC models frequently fail to converge properly (Figure 2).

The convergence problems are worse for the FIRC model two, because it requires more data to fit the running variable coefficients as random effects. Even when the sample size is increased to 300 total schools the FIRC Two model fails to converge 70% of the time and when the average site size is 250 observations it still fails to converge 63% of the time. Convergence does improve for

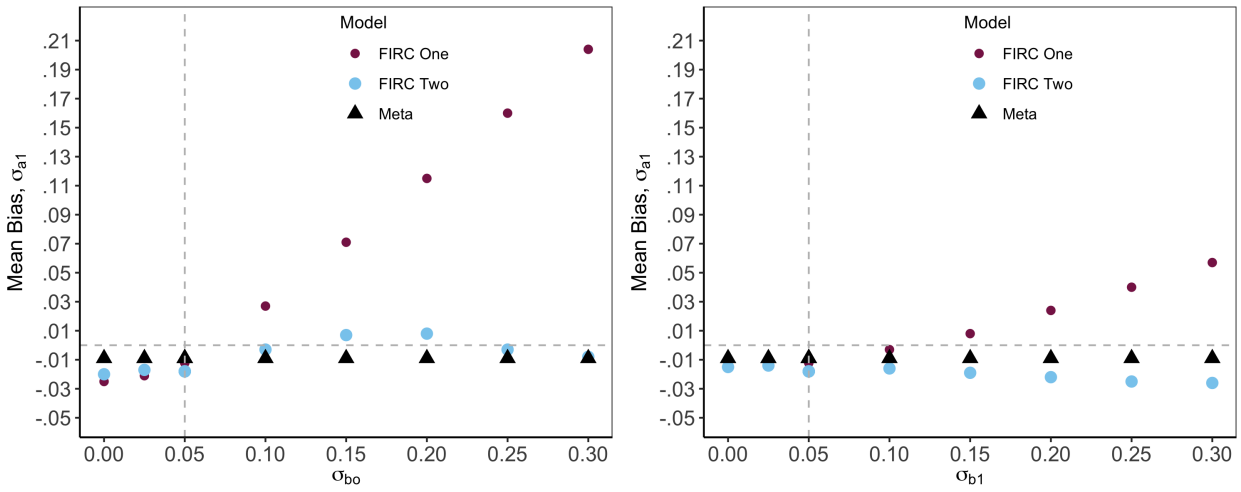
the FIRC One model as the sample size increases. Once the total number of sites reaches 250 or the average site size reaches 150 the model fails to converge less than 10% of the time. However, when the sample is smaller, convergence also presents a large problem for the FIRC One model.



(a) (b)

Figure 2: The rate at which the FIRC models produce a singular estimate of the treatment effect standard deviation. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right).

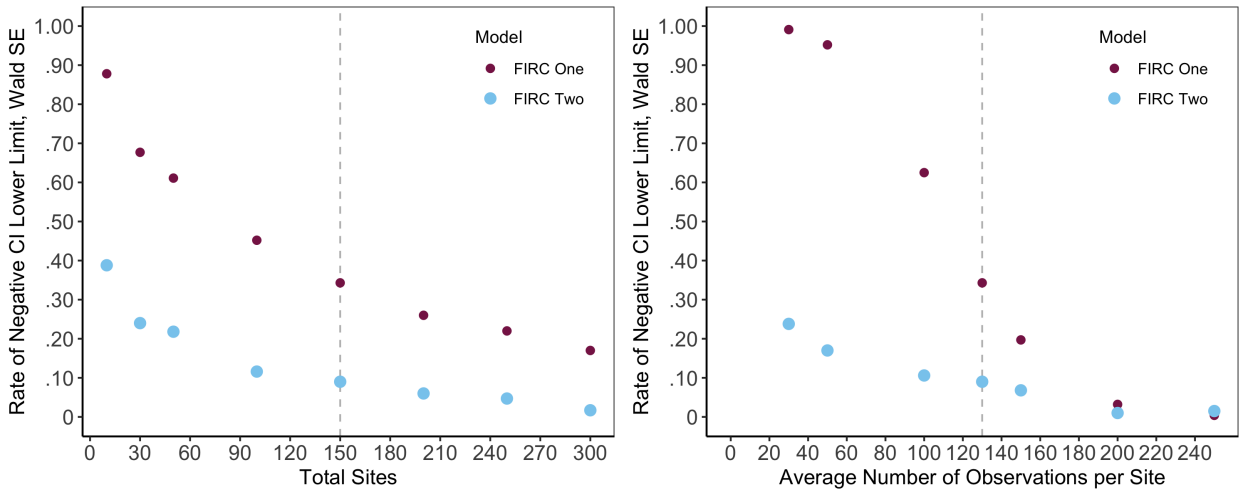
The FIRC One model's more consistent convergence comes at a cost. The pooled estimates of the running variable coefficients make the model misspecified. This modeling treats the running variable coefficients as though their cross-site variance is zero. If there is significant cross-site variance in the running variable coefficients, the model interprets it as variance in the cross-site treatment effect, and the cross-site treatment effect variance estimate becomes biased (Figure 3).



(a) (b)

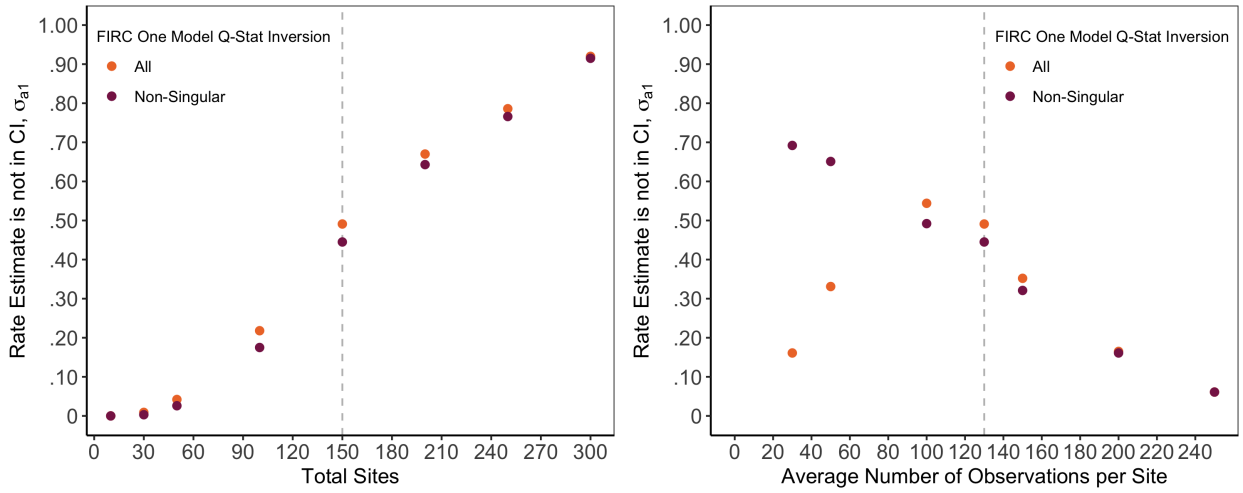
Figure 3: The mean bias in the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), fixed intercepts random coefficients with random running variable coefficients (FIRC Two), and random effects meta-analysis (Meta), regression discontinuity models. Both FIRC models exclude the instances of singular fit. In the left panel, the standard deviation of b_1 is fixed at .05, and b_0 is varied. In the right panel the standard deviation of b_0 is fixed at .05, and b_1 is varied. In both panels, the dotted line is at the baseline parameter value of .05

Neither the Wald standard errors nor the Q-statistic inversion method work well for calculating the confidence intervals of the cross-site treatment effect standard deviation estimates from the FIRC models. Figure 4 shows that the Wald standard errors frequently have confidence intervals that go below zero, which is inaccurate because standard deviation values can not be negative. Estimating confidence intervals using Q-statistic inversion has a different problem. Recall that Q-statistic inversion uses a modified version of the FIRC model estimated using OLS and not a multi-level model. In the case of the FIRC RDD model, using this new OLS model creates a situation where the actual parameter estimate is frequently not in the confidence interval (Figure 5). This problem occurs regardless of whether instances where the model has a singular fit are including or excluded. Given these problems with Wald standard error and the Q-statistic inversion method, we use the profile method for getting confidence intervals for the FIRC models in the rest of this paper.



(a) (b)

Figure 4: The rate at which the lower limit of the Wald confidence interval goes below zero for the two FIRC models. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right).

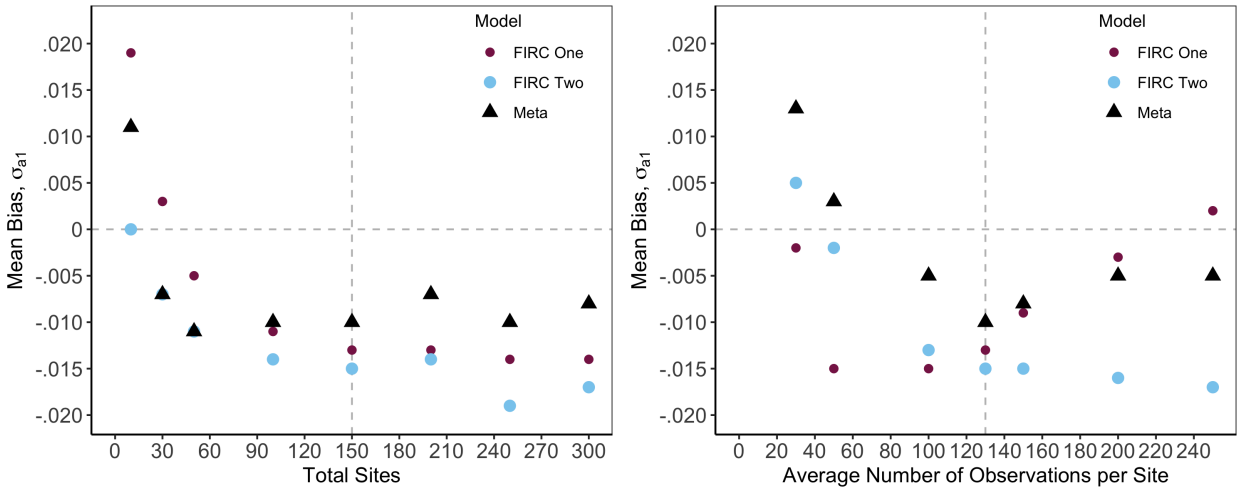


(a) (b)

Figure 5: The rate at which the FIRC One model treatment effect standard deviation estimate is not in the confidence intervals obtained from Q-statistic inversion. Results from the FIRC Two model are not included for clarity of presentation. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right)

Evaluating The Random Effects Meta-Analysis Model

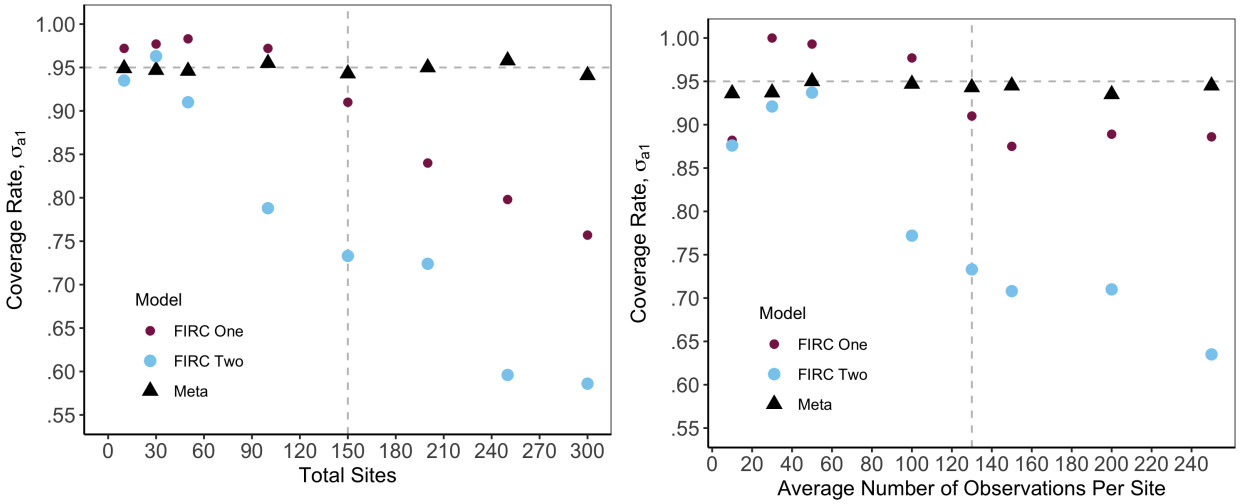
Our proposed random effects meta-analysis model does not suffer from the same problems as the FIRC model. Unlike the FIRC One model, cross-site variance in the running variable coefficients does not induce bias in the cross-site treatment effect estimate (Figure 3). In Figure 6, we present the mean bias for the three model estimates across a range of sample sizes using the baseline parameter values. There is a small amount of downward bias in the parameter estimates in all three models, including the random effects meta-analysis model. However, when the total number of sites is great than ten or the average number of observations is greater than 50, the random effects meta-analysis model has less bias than the FIRC Two model. The random effects meta-analysis also has less bias than the FIRC One model as long as the total number of sites is at least 100. The random effects meta-analysis model has less bias than the FIRC One model for most of the site sizes we tested, however as the average site size gets very big, the FIRC One model has less bias. This is because the downward bias from the multi-level model is canceled out by the upward bias from the model misspecification.



(a) (b)

Figure 6: The mean bias in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity models. Both FIRC models exclude the instances of singular fit. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right).

Despite the small amount of bias in the random effects meta-analysis estimates of the cross-site treatment effect standard deviation, the bias is well within the confidence interval and does not impact the coverage rate. The random-effects meta-analysis estimate's coverage rate is about 95% across sample sizes, which is not true for the FIRC models (Figure 7). The FIRC Two model only has a coverage rate near 95% when the sample sizes are small, and therefore when the model is least likely to run successfully. The FIRC One model coverage is closer to 95%, but when there are 200 total sites, the FIRC One coverage is only 84% and continues to drop as total sites increases.

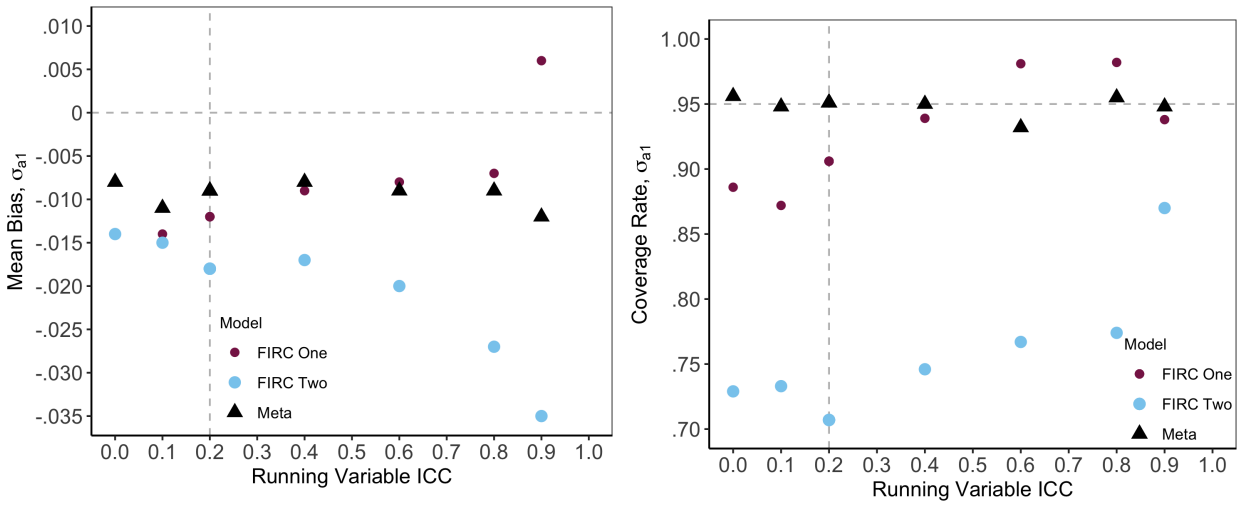


(a)

Figure 7: The coverage rate of the cross-site treatment standard deviation confidence intervals for the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity models. Both FIRC models exclude the instances of singular fit and use the profile method for obtaining confidence intervals. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right).

The mean bias and coverage results remain similar when we vary the ICC of the running variable instead of the sample size (Figure 8). The mean bias in random effects meta-analysis estimate of the cross-site treatment effect standard deviation is small, and the coverage of the confidence interval is close to 95% across ICC values. This is also mostly true for the FIRC One model, except that the coverage is low for small ICC values. However, the mean bias in the FIRC Two model estimate of the cross-site treatment effect standard deviation is affected by increasing

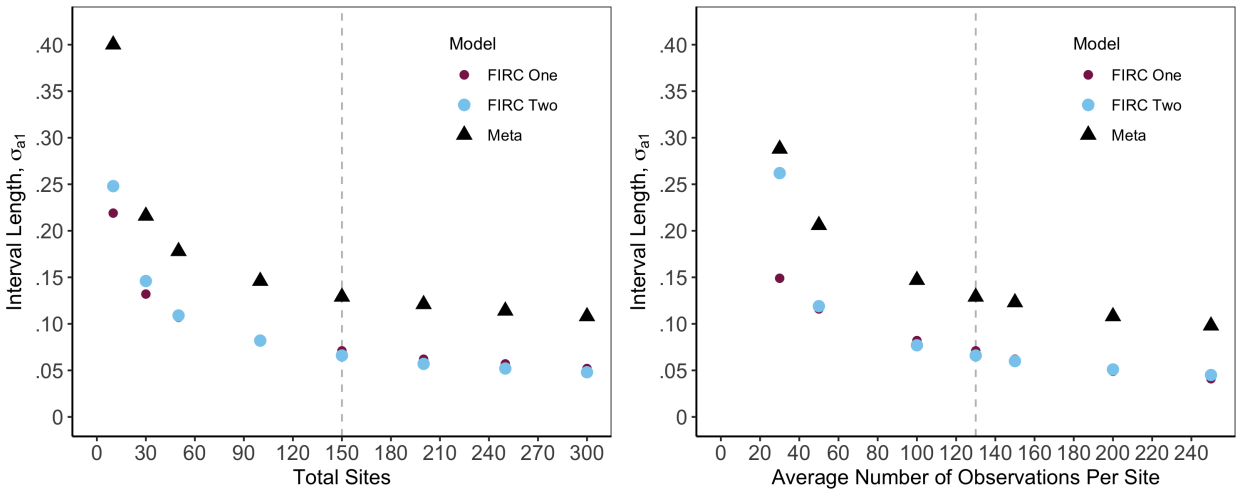
the ICC and get much worse for higher ICC values.



(a)

Figure 8: The accuracy of the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), fixed intercepts random coefficients with random running variable coefficients (FIRC Two), and random effects meta-analysis (Meta), regression discontinuity models. Both FIRC models exclude the instances of singular fit and use the profile method for obtaining confidence intervals. The right panel contains the mean bias in the estimate as the running variable ICC is varied, and the left panel contains the coverage rate of the confidence intervals as the ICC is varied.

There is a metric by which the meta-analysis model consistently performs worse than the FIRC models. The confidence intervals for the meta-analysis model are longer than the confidence intervals for the FIRC model (Figure 9). We argue that this does not justify using either of the FIRC models because of the other demonstrated problems with those models. However, unlike bias or coverage, interval length is a metric that can be assessed ex-post when the model is being used in practice on non-simulated data. This means a researcher trying to decide between models might reasonably conclude one of the FIRC models is better if they were not aware of these FIRC models' other problems.



(a) (b)

Figure 9: The length of the confidence interval of the cross-site treatment standard deviation confidence intervals for the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity models. Both FIRC models exclude the instances of singular fit and use the profile method for obtaining confidence intervals. In the left panel, the average number of observations per site is fixed at 130, and the total number of sites is varied. In the right panel, the total number of sites is fixed at 150, and the average number of observations is varied. In both panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 130 average observations per site (right).

Massachusetts Education Proficiency Plan Example

Background

For the last 15 years, students in Massachusetts have been required to pass the 10th grade MCAS exams in English Language Arts (ELA) and Mathematics in order to graduate. Students pass the MCAS if they achieve at least the minimum score to be designated "Needs Improvement". In 2006, the law in Massachusetts was changed and, starting with the 2010 graduating cohort, students who scored high enough in ELA or Math to be designated as "Needs Improvement" but not high enough to be "Proficient" now must complete an Education Proficiency Plan (EPP) in their nonproficient subject.

The EPP policy was established by statute at the state level but is implemented by individual high schools. High schools across Massachusetts have considerable latitude in how they implement EPPs for their students. High schools can require students to demonstrate proficiency

by taking a special proficiency exam, passing courses in the relevant area(s) in their junior and senior year, or a combination of the two. In the end, final proficiency is certified locally by a student's own principal. High schools have the most latitude in how math EPPs are implemented. Massachusetts has no state-wide rule regarding how much math and ELA high schools must require for graduation. In practice, however, all Massachusetts high schools require four years of ELA to graduate, but high schools range from requiring 2 to 4 years of math to graduate.

The EPP policy's adoption was part of a larger push from the Massachusetts Board of Elementary and Secondary Education to increase the number of high school students who completed a math course in their senior year. Therefore, one relevant question about the EPP is whether students who were required to complete a math EPP were more likely to complete a math course their senior year. We answer this question using an RDD, with the raw 10th grade math MCAS score as the running variable and whether a student completes a math course two years after the MCAS exam as the outcome variables. Massachusetts started collecting course taking data in 2011. For each graduating cohort from 2011 to 2016, we separately estimate the effect of being required to complete a math EPP on the probability of completing a math course two years after taking the MCAS, which we use as a proxy for completing a math class in a student's senior year.

When thinking about the EPP policy, the average treatment effect is not the only quantity of interest. Given that EPPs were administered at the high school level, it is important to understand how much between high school variation there was in the treatment effect. The state of Massachusetts is not only concerned with how the EPP policy affects the average student but the whole distribution of effects. Even if a policy helps students on average, it is of policy interest to know whether it also harms a substantial number of students. Understanding treatment variation also provides information on how to target implementation support. If the treatment effect variance is low, it makes sense to target supports broadly, and if the treatment effect variance is high, it makes sense to focus supports on the schools where the policy is working poorly. In this example, we, therefore, also estimate the treatment effect variance across high schools. In addition to the main analysis, we also estimate treatment effects and treatment effect

standard deviations separately for schools that required four years of math and for schools that required less than four years of math.

Results

Students required to complete a math EPP were more likely to complete a math class their senior year than students who were not required to complete a math EPP (Figure 10). Students in the 2011 cohort bound by the math EPP were seven percentage points more likely to complete a math class their senior year than those not bound by the EPP. We see that the math EPP's effect declined over time and that by the 2016 cohort, students required to complete the EPP were only three percentage points more likely to two years after taking the MCAS.

The math EPP policy corresponded with other policies intended to increase the number of high schools that required four years of math to graduate and, ultimately, the number of high school seniors who took and passed math classes. One reason we see the effect of the math EPP declining across cohorts is more high schools required four years of math to graduate, and so students not bound by the math EPP were also required to take math their senior year. Overall the baseline percentage of students completing math classes their senior year was also increasing across these cohorts.

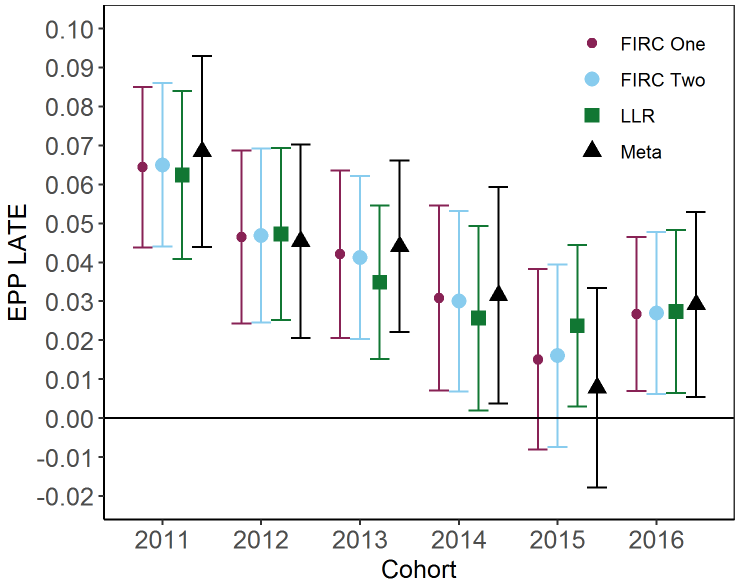
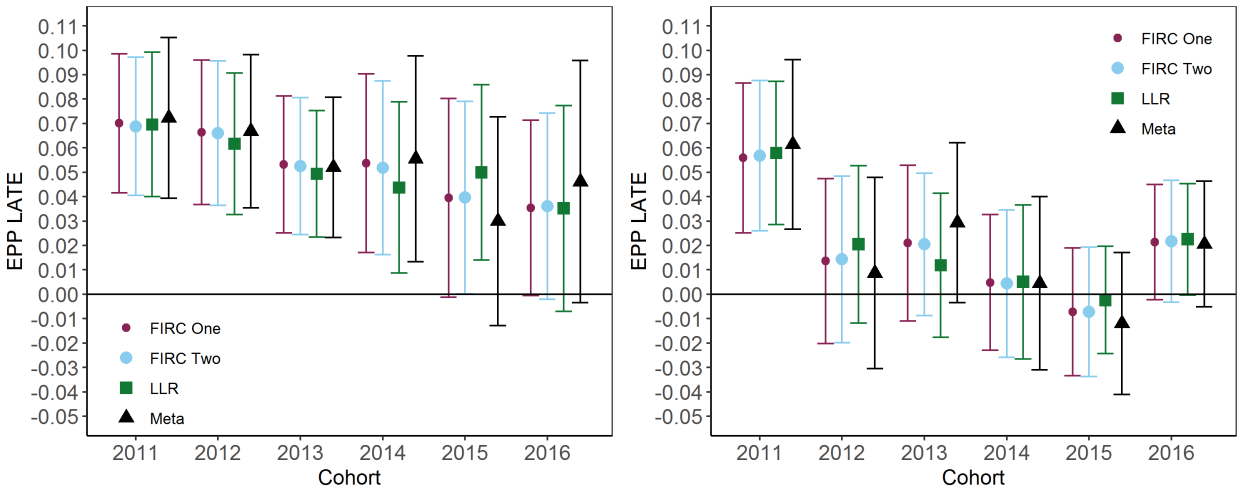


Figure 10: The local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts.

When we split the sample by whether a high school required four years of math to graduate high school, the EPP effect is more consistent across cohorts (Figure 11). In high schools that require less than four years of math, the EPP effect goes from about seven percentage points in the 2011 cohort to about five percentage points in the 2016 cohort. However, the estimates get increasingly noisy over time as the sample of schools that do not require four years of math gets smaller. In high schools that require four years of math, the EPP effect is about six percentage points in the 2011 cohort, but for all the other cohorts, it is consistently near zero and not statistically significant.



(a) High Schools Requiring ≤ 4 Years of Math to Graduate (b) High Schools Requiring 4 Years of Math to Graduate

Figure 11: The local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts by whether the high school requires four years of math to graduate.

Finally, the model choice does not significantly affect our estimates of the average treatment effect. Across cohorts and samples, all four models produce similar average treatment effect point estimates and confidence intervals. As with the simulations, the average treatment effect estimate is robust to different assumptions about pooling the treatment coefficient or the running variable coefficients. Also consistent with the simulations, the standard local linear model that is generally used to estimate average treatment effects in multisite RDD's doesn't produce different estimates than the other models.

While the math EPP's average effect fell across cohorts, there is no consistent pattern in the cross-high school treatment effect variance. In the 2012, 2013, and 2015 cohorts, we can not reject the null hypothesis that cross-high school treatment effect standard deviation is zero, and in the 2011, 2014, and 2016 cohorts, we can reject the null hypothesis. Across cohorts, the standard deviation point estimate bounces between 10 percentage points for the 2014 cohort and four percentage points for the 2013 cohort (Figure 12). If we assume that the treatment effect is normally distributed across schools, then this amount of cross-high school variation implies that for the 2011, 2014, and 2016 cohorts in many Massachusetts high schools, the math EPP had the opposite of the desired effect and reduced the likelihood that a student completed a math class

their senior year. These results could be driven by either variation in the treatment implementation or in the control group behavior. However since non-EPP students were more consistently taking math their senior year over this time period, these results imply that the EPP policy implementation was not getting more consistent across high schools as the policy got older.

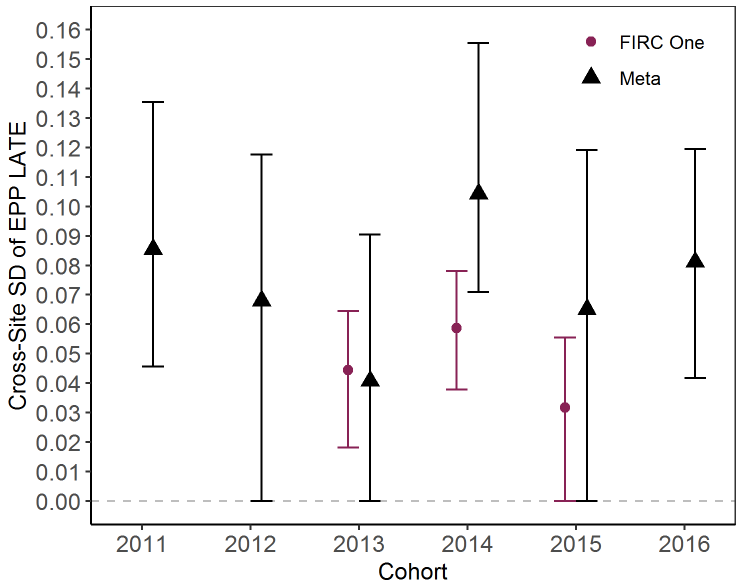
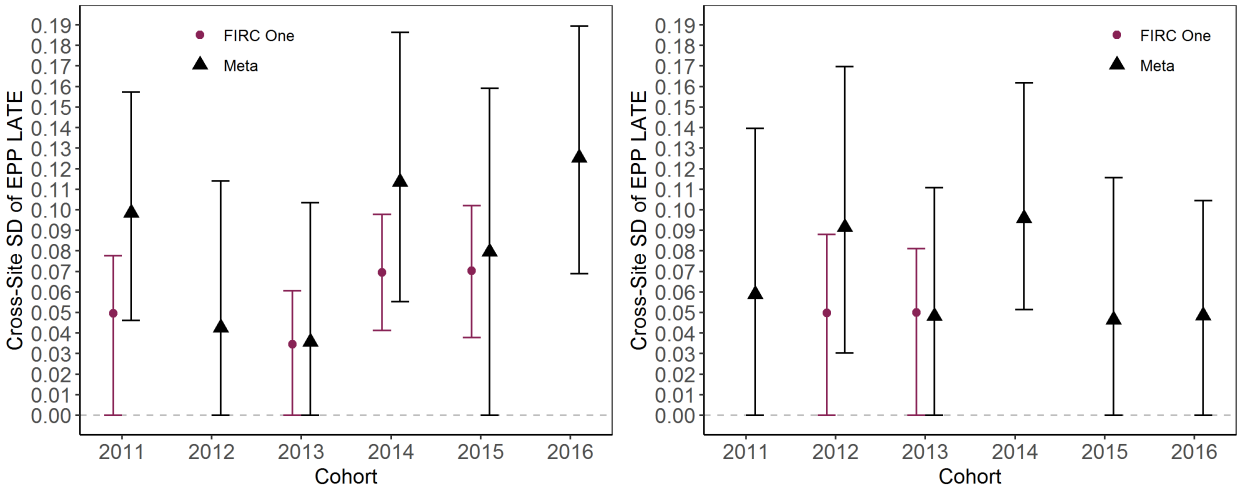


Figure 12: The cross-high school standard deviation of the local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts.

RDD papers often capture treatment variation by looking at treatment effect heterogeneity by observable characteristics, as in our analysis in Figure 11. However, by estimating the cross-site treatment effect standard deviation, we can show that high school graduation requirements explain only a little of the variation (Figure 13). We can see this in two ways. First, there is still significant variation within each group of schools. For schools that do not require four years of math, the cross-site standard deviation is statistically significant in three cohorts, and for schools that do require four years of math, the cross-site standard deviation is statistically significant in two cohorts. Therefore the overall cross-site treatment effect variation is not fully explained by differences in the treatment effect across the two groups of high schools. Second, because there is still treatment effect variance in the schools requiring four years of math, we can see that the graduation requirements are not the only thing driving treatment effect variation. Since even within schools where the EPP should not be a binding constraint on students' senior year math

coursework decisions, we see variation in the EPP effect. Overall, we can therefore see the value of quantifying the variance on top of just estimating heterogeneity by observable characteristics.



(a) High Schools Requiring ≤ 4 Years of Math to Graduate (b) High Schools Requiring 4 Years of Math to Graduate

Figure 13: The cross-site standard deviation of the local average treatment effect of the Education Proficiency Plans on the probability of completing a math course senior year across cohorts by whether the high school requires four years of math to graduate.

The EPP example also highlights the concern with the FIRC models seen in the simulations. The FIRC Two model provided a singular fit in all eighteen of our models and so is not picture in the figures. The FIRC One model produced a non-singular fit in only half of the models we ran. The FIRC model confidence intervals are also tighter. A reasonable researcher might look at the treatment effect standard deviation estimates from these two models and pick the FIRC One model. However, the FIRC One model estimates of the cross-site standard deviation are lower than the random effects meta-analysis estimates. Consistent with Figure 6, we should expect this because the FIRC One estimates are more biased than meta-analysis estimates and are worse estimates of the cross-site treatment effect standard deviation. Therefore, the FIRC One model serves as an attractive nuisance to researchers, which they should be wary of.

Conclusion

Understanding treatment effect variation is an important part of policy evaluation. Within

RCTs, there are increasingly standard methods for estimating treatment effect variation. In this paper, we show that adapting these methods to the RDD setting is complicated, and methods that work in the context of RCTs do not work with RDDs. We develop and evaluate a method for estimating cross-site treatment effect variance within an RDD based in meta-analysis. We also evaluate and expand two FIRC models used in prior research. We demonstrate the FIRC models work poorly in the RDD context, but that our random effects meta-analysis model performs well across various conditions.

We then apply these methods for estimating cross-site treatment effect variation to a practical policy problem. We evaluate the effect of Massachusett's Education Proficiency Plans on senior year math completion rates. We find that the Education Proficiency Plans did increase senior year math completion rates, but we also find that there was substantial variation across high schools in this effect. This implies an opportunity for the state to improve the policy's effectiveness by targeting schools where the policy is less effective with increased implementation supports.

References

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177–188.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), 615–635.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, *48*(2), 281–355.

McEachin, A., Domina, T., & Penner, A. (2020). Heterogeneous effects of early algebra across california middle schools. *Journal of Policy Analysis and Management*, *39*(3), 772–800.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475–499.

Raudenbush, S. W., Reardon, S. F., & Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of research on Educational Effectiveness*, *5*(3), 303–332.

Tipton, E. (2014). How generalizable is your experiment? an index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in medicine*, *10*(11), 1665–1677.