ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-422

Characterizing Cross-Site Variation in Local Average Treatment Effects in Multisite Regression Discontinuity Design Contexts with an Application to Massachusetts High School Exit Exam

Sophie Litschwartz
Harvard University

Luke Miratrix
Harvard University

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). This research rectifies this gap in the literature by systematically exploring and evaluating methods for estimating the cross-site treatment effect variance in multisite RDDs. Specifically, we formalize a fixed intercepts/random coefficients (FIRC) RDD model and develop a random effects meta-analysis (Meta) RDD model for estimating cross-site treatment effect variance. We find that a restricted FIRC model works best when the running variables' relationship to the outcome is stable across sites but can be biased otherwise. In those instances, we recommend using either the unrestricted FIRC model or the meta-analysis model; with the unrestricted FIRC model generally performing better when the average number of in-bandwidth observations is less than 120 and the meta-analysis model performing better when the average number of in-bandwidth observations is above 120. We apply our models 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 on average across sites, but that the impact varied enough such that a third of schools could have had a negative impact.

VERSION: June 2021

Suggested citation: Litschwartz, Sophie, and Luke Miratrix. (2021). Characterizing Cross-Site Variation in Local Average Treatment Effects in Multisite Regression Discontinuity Design Contexts with an Application to Massachusetts High School Exit Exam. (EdWorkingPaper: 21-422). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/98tk-5w98

Characterizing Cross-Site Variation in Local Average Treatment Effects in Multisite Regression Discontinuity Design Contexts with an Application to Massachusetts High School Exit Exam

Sophie Litschwartz
Harvard Graduate School of Education

Luke Miratrix
Harvard Graduate School of Education

June 10, 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). This research rectifies this gap in the literature by systematically exploring and evaluating methods for estimating the cross-site treatment effect variance in multisite RDDs. Specifically, we formalize a fixed intercepts/random coefficients (FIRC) RDD model and develop a random effects meta-analysis (Meta) RDD model for estimating cross-site treatment effect variance. We find that a restricted FIRC model works best when the running variables' relationship to the outcome is stable across sites but can be biased otherwise. In those instances, we recommend using either the unrestricted FIRC model or the meta-analysis model; with the unrestricted FIRC model generally performing better when the average number of in-bandwidth observations is less than 120 and the meta-analysis model performing better when the average number of in-bandwidth observations is above 120. We apply our models 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 on average across sites, but that the impact varied enough such that a third of schools could have had a negative impact.

Keywords: *cross-site variation; regression discontinuity designs; high school exit exams*

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 setup 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 approximately three 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; Kling, Liebman, & Katz, 2007; Raudenbush, Reardon, & Nomi, 2012).

One measure of treatment effect variation in multisite studies is the cross-site treatment effect variance. There are standard methods for estimating cross-site treatment effect variance in multisite randomized experiments, but no such methods are 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.

Despite there being no standard way to estimate the cross-site treatment effect variance in multisite RDDs, a few studies have adapted experimental models for estimating cross-site treatment effect variance and used them to estimate cross-site treatment effect variance in a multisite RDD. Raudenbush et al. (2012) in a study on statistical methods for multisite instrumental variable analysis, estimate the cross-site variance of the first stage of a multisite RDD evaluation of double dose algebra in Chicago schools. Raudenbush et al. (2012) estimate the cross-site variance using a multi-level model where the outcome, taking a double dose algebra course, is estimated as function an intercept, treatment status, a linear running variable term, and quadratic running variable term and where all coefficients, except the coefficient on the quadratic running variable term, are estimated using random effects. McEachin, Domina, and Penner (2020) estimate the cross-site treatment effect variance in a multisite study of early algebra in California middle schools. McEachin et al. (2020) estimate the cross-site treatment effect variance using a fixed intercepts random coefficient model (FIRC), where the outcome of interest is modeled as a function of a school-year fixed effect, treatment, a linear running variable term, and a linear treatment running variable interaction; in this model, only the treatment is estimated with a random coefficient, and all other coefficients are fixed across the sample. Shapiro (2020) estimates the cross-site treatment effect variance in a multisite RDD study of the effect of age at enrollment on special education placement in Michigan. Shapiro (2020) also estimates the cross-site treatment effect variance using a FIRC model, where special education placement is modeled as a function of district level intercept, treatment, a linear running variable term, and a vector of relevant covariates. As with McEachin et al. (2020), the treatment coefficient is estimated as a random effect, and all the other coefficients are fixed across the sample.

Each of these prior studies is adapting the RCT methods for estimating cross-site treatment effect variance differently, and none of these methods have ever been evaluated in an RDD context. In the first part of this paper, we formalize two methods for estimating cross-site treatment effect variance in a multisite RDD and use simulation to evaluate under which conditions each should be used. In line with the prior multisite RDD studies, which have estimated cross-site treatment effect variance, our first model is a hybrid of the local linear multisite RDD model and the multisite RCT FIRC model. We test two versions of the FIRC model, a restricted FIRC where we estimate the running variable coefficients as fixed across sites and an unrestricted FIRC model where we fit random effects on the running variable coefficients. We also compare maximum likelihood estimation to restricted maximum likelihood estimation and evaluate three potential methods for estimating confidence intervals: Wald standard errors, Q-statistics inversion, and profiled confidence intervals.

The second method for estimating cross-site treatment effect variance is unique to this study and based on random effects meta-analysis. In this method, 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. This estimation procedure allows each site to have varying and unconstrained relationships between the running and outcome variables. 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 find a key driver of performance between methods is whether the outcome running variable relationship is stable across sites. If it is, the restricted FIRC model works well. However, when there is variance in the running variable coefficients, the restricted model treats this running variable coefficient variance as variance in the treatment effect, and the cross-site treatment effect estimate is upwardly biased. To select between options, we recommend using a model selection criteria (e.g., AIC) to test whether the restricted or unrestricted FIRC model produces a better fit. When there is variance in the running variable coefficient, the random effects meta-analysis model has less bias and error than the unrestricted FIRC model when the average number of

in-bandwidth observations per site is large, generally when the average number of in-bandwidth observations above 120. Otherwise, the unrestricted FIRC model has less error than the random-effect meta analysis model and, except for very small sample sizes, less bias than the random-effect meta analysis model.

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 multi-site RDD models 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. Controlling for graduation requirements, we consistently find statistically significant cross-site treatment effect variation amongst high schools that did not require four years of math and, while there is less treatment effect variation amongst high schools requiring four years of math to graduate, we still

find statistically significant cross-site treatment effect variation in those schools in three of the six cohorts we examined. Therefore we can conclude there were meaningful differences in program implementation across schools beyond differences in course requirements.

Analytical Models

A linear RDD model frequently takes the following form:

$$\begin{gathered} Y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_2(Score_{ij} - Score_c) + \beta_3 T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\sim} N[0, \sigma_y^2] \end{gathered} \tag{1}$$

where Y_{ij} is the outcome of interest, $Score_{ij}$ 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 analysis β_0 is the intercept, β_1 is the local average treatment effect (LATE), β_2 is the relationship between the running variable and the outcome, and β_3 is how much the treatment effect varies by the running variable.

In a multisite study, for each of these four coefficients, we can choose to estimate these coefficients by pooling, partially pooling, or fully unpooling data across sites. A pooled coefficient is estimated by fully combining data from across all sites. This modeling assumes no cross-site variance in the coefficient value, and one coefficient value is estimated jointly across sites. A partially pooled coefficient is assumed to be normally distributed across the sites and estimated as a site-level random effect using a multi-level model. A fully unpooled coefficient makes no assumption about how the coefficient is distributed across sites. A separate coefficient value is estimated for each site, using only data from that site.

Currently, multisite RDD studies generally estimate causal effects using a local linear regression with an unpooled site level intercept and then with all other coefficients fully pooled (Hahn, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008; Gelman & Imbens, 2019). In this model, 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.

We present three potential models for estimating cross-site effect variance in a multi-site RDD (Table 1). The first two models are fixed intercepts random coefficient (FIRC) models. A FIRC model contains unpooled intercepts, the fixed intercepts, and a partially pooled LATE coefficient, the random coefficient. The first FIRC model (FIRC One) we evaluate is restricted, and we estimate the two running variable coefficients as fully pooled across sites. The second FIRC model (FIRC Two) is unrestricted, and we estimate the two running variable coefficients as partially pooled across sites.

The final model (Meta) 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 using only data from that site. Therefore, in this model, each of the four regression coefficients is fully unpooled across sites.

Model	Unpool	Partially Pool	Pool	Treatment Effect Heterogeneity Allowed
Local Linear Regression (LLR)	β_0		β_1, β_2, β_3	No
Random Effects Meta-Analysis (Meta)	β_0, β_1, β_2, β_3			Yes
Fixed Intercepts Random Coefficient One (FIRC One)	β_0	β_1	β_2, β_3	Yes
Fixed Intercepts Random Coefficient Two (FIRC Two)	β_0	β_1, β_2, β_3		Yes

Table 1

The Fixed Intercepts Random Coefficient Models:

We evaluate two FIRC models, a restricted model with pooled coefficients on the running variable terms and an unrestricted model with partially pooled coefficients on the running variable terms. The restricted model is more parsimonious but implicitly assumes no cross-site variance in the running variable coefficients. In Appendix A, we also evaluate a partially restricted FIRC model with a partially pooled β_2 and a pooled β_3, that lies between our two FIRC models explored here.

These two FIRC models are as follows:

FIRC One (restricted)

Level One - Observation:

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

Level Two - Site:

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

FIRC Two (unrestricted)

Level One - Observation:

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

Level Two - Site:

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

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. We estimate both the models using restricted maximum likelihood (REML), however, we also present results using maximum likelihood in Appendix B.

The Random-Effects Meta Analysis Model:

The first step of the Meta model is estimating the following regression model in each site:

$$\begin{gathered} Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}(Score_{ij} - Score_c) + \beta_{3j}T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\sim} N[0, \sigma^2_{yj}] \end{gathered} \tag{4}$$

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

$$\begin{aligned} VCov(\widehat{\beta_j}) &= (X_j' X_j)^{-1} \widehat{\sigma_j}^2 \\ \widehat{\sigma_j^2} &= \frac{1}{n_j - 3 - 1} \sum (Y_{ij} - \widehat{Y_{ij}})^2 \end{aligned} \tag{5}$$

where X_j is the data matrix of dependent variables (i.e. treatment status, the running variable value, and the treatment running variable interaction) 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:

$$\begin{aligned} VCov(\widehat{\beta_j}) &= (X_j' X_j)^{-1} \overline{\widehat{\sigma_j}^2} \\ \overline{\widehat{\sigma_j^2}} &= \frac{1}{\sum n_j} \sum (n_j \hat{\sigma_j}^2) \end{aligned} \tag{6}$$

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. Therefore the overall local average treatment effect is estimated as follows:

$$\widehat{\beta_1} = \frac{\sum \widehat{\beta_{1j}} w_j}{\sum w_j},\ \widehat{SE_{\beta_1}} = \sqrt{\frac{1}{(\sum w_j)}} \tag{7}$$

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 the DerSimonian-Laird (DL) methods of moments estimator:

$$\begin{aligned} \widehat{\sigma^2_{\beta 1}} &= max(0, \frac{Q - J - 1}{\sum \widehat{SE^{-2}_{\beta_{1j}}} - \frac{\sum \widehat{SE^{-4}_{\beta_{1j}}}}{\sum \widehat{SE^{-2}_{\beta 1j}}}}) \\ Q &= \sum W_j (\widehat{\beta_{1j}} - \frac{\sum W_j \widehat{\beta_{1j}}}{\sum W_j})^2, where\ W_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}}} \end{aligned} \tag{8}$$

where $\widehat{SE_{\beta_{1j}}}$ is the estimated site-level standard error for β_{1j} from Equation 6 and J is the total number of sites.

Q-Statistic Inversion Confidence Intervals:

For all three models, the confidence interval for the cross-site treatment effect variance is estimated using Q-statistic inversion. For the FIRC models we test two additional methods for obtaining confidence intervals (Appendix C): Wald standard errors and profiled confidence intervals. Q-statistic inversion intervals perform better than these two methods and therefore we use that method in our analysis.

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

$$Q(\tau^2) = \sum_{j=1}^{J} \frac{(\widehat{\beta_{1j}} - \frac{\sum W_j \widehat{\beta_{1j}}}{\sum W_j})^2}{\widehat{SE^2_{\beta_{1j}}} + \tau^2}, where\ W_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}}} \tag{9}$$

The Q-statistic $Q(\tau^2)$, is similar to the Q in the DL estimator, but $Q(\tau^2)$ includes the treatment effect variance (τ^2) in the denominator (Higgins et al., 2009). Under the true τ^2, 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 the meta-analysis model $\widehat{\beta_{1j}}$ and $\widehat{SE^2_{\beta_{1j}}}$ are taken directly from Equations 4 and 6. In the FIRC models these quantities are not estimated from the original multi-level model but a new OLS regression model. A new OLS model is required because the site level treatment effects are designed to be individually interpretable and are over shrunk for use as distribution level estimates. These new OLS models take the following forms:

FIRC One (restricted)

$$\begin{gathered} Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_2 Score_{ij} + \beta_3 T_{ij} * Score_{ij} + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_y) \end{gathered} \tag{10}$$

FIRC Two (unrestricted)

$$\begin{gathered} Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j} Score_{ij} + \beta_{3j} T_{ij} * Score_{ij} + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_y) \end{gathered} \tag{11}$$

The FIRC Two confidence intervals from this process work out to be the same in expectation as the Meta confidence intervals, but may be slightly different for any given set of data.

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, given a vector of site characteristics (i.e., a_{0j}, a_{1j}, b_{0j}, b_{1j}, and r_j) data for observation i in site j is generated as follows:

$$\begin{aligned} Y_{0ij} &= a_{0j} + b_{0j} Score_{ij} + \epsilon_{ij} \\ Y_{1ij} &= Y_{0j} + a_{1j} + b_{1j} Score_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma^2_\epsilon) \\ Score_{ij} &= \mu_{Score} + r_j + \pi_{ij} \\ \pi_{ij} &\sim N(0, 1 - ICC_{Score}) \end{aligned} \tag{12}$$

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 a site-level offset on the running variable, and ICC_{Score} is the inter-class correlation of the running variable. In this model, the cutpoint $Score_c$ 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}. Each site is also defined as having n_j observations and the residual error variance, σ^2_ϵ, is assumed to be fixed across sites and observations

Under this data generating process, all coefficients are site specific and the site level parameters are generated as follows:

$$\begin{aligned} a_{0j} \sim N(\mu_{a0}, \sigma^2_{a0}), b_{0j} \sim N(\mu_{b0}, \sigma^2_{b0}), a_{1j} \sim N(\mu_{a1}, \sigma^2_{a1}), b_{1j} \sim N(\mu_{b1}, \sigma^2_{b1}) \\ r_j \sim N(0, ICC_{Score}) \\ n_j \sim Pois(\mu_n) \end{aligned} \tag{13}$$

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 independent from each

other.

The simulation input parameters are then the 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} are 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, the running variable grand mean (μ_{Score}) to 1, and the running variable ICC (ICC_{Score}) to a value of .2.

We run three main groups of simulations. The first group of simulations is under the conditions where the FIRC One model is correctly specified, and there is no cross-site variance in the running variable coefficients (i.e., σ_{b0} and σ_{b1} are fixed at zero). Across these simulations, we vary the average observations per school (μ_n) from 10 to 350, holding the total number of schools fixed at 150, and we vary the total schools (J) from 10 to 300, holding the average observations per school fixed at 130. In the second group of simulations, we set σ_{b0} and σ_{b1} equal to .05; this makes the FIRC One model misspecified because there is cross-site variance in the running variable coefficients. In this group of simulations, we run all sample size combinations as we vary the average observations per school from 10 to 350 and the total schools from 10 to 300. In the third group of simulations, the average number of observations per school is fixed at 130, and the total number of schools is fixed at 150. We then run simulations where σ_{b1} is fixed at .05 and the standard deviation of the control running variable coefficient (σ_{b0}) is varied from 0 to .3 and simulations where σ_{b0} is fixed at .05 and the standard deviation of the treatment running variable coefficient (σ_{b1}) is varied from 0 to .3.

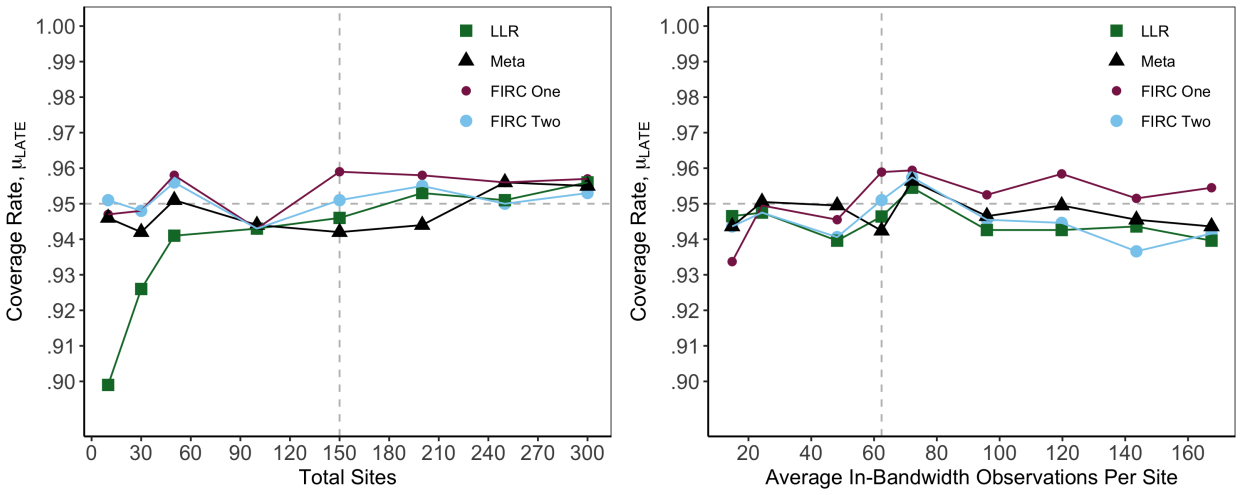
Finally, in Appendix D we run a set of simulations where the interclass correlation of the

running variable (ICC_{Score}) is varied from 0 to .9 to demonstrate the results are not sensitive to changes in the interclass correlation of the running variable. Additionally, while the overall average number of observations per site is directly manipulated in our simulations, we report results using the average number of in-bandwidth observations per school because the number of in-bandwidth observations more directly affects the simulation results.

Simulation Results

Estimate of the Treatment Effect Mean

The LLR model, the Meta model, and both the FIRC models produce reasonable estimates of the local average treatment effect. Using the benchmark parameter values and assuming σ_{b0} and σ_{b1} equal to .05, 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 different site sizes and 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, and the total number of sites is varied. In the right panel, the total number of sites is fixed, 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 62 average in-bandwidth observations per site (right).

The FIRC Models

The FIRC One model assumes no variance in the running variable coefficients. Phrased another way, this model assumes the relationship between the running variable and the outcome is constant across sites and that the treatment has the same effect on the relationship between the running variable and the outcome in every site. When this assumption holds (i.e. in simulations where $\sigma_{b0} = 0$ and $\sigma_{b1} = 0$), then the FIRC One model produces the best estimates. Figure 2 shows the mean bias and root mean squared error (RMSE) for the three different models when there is no variance in the running variable coefficients. All the model estimators of the cross-site treatment effect standard deviation have some bias in their estimates across a range of sample sizes; however, the bias is consistently the smallest for the FIRC One model. The FIRC One model estimate also consistently has the smallest RMSE.

However, when there is cross-site variance in the running variable coefficients, the FIRC One model estimates become upwardly biased (Figure 3). In an RDD, there is no common

support for the running variable across treatment and control; this makes the RDD model particularly sensitive to misspecification in the running variable modeling. Therefore, the bias occurs because the FIRC One model interprets cross-site variance in the running variable coefficients as cross-site variance in the treatment effect. The upward bias also worsens as the amount of running variable variance increases (Figure 4). In Figure 4 more bias is induced by the variance in the running variable coefficient than by variance in the running variable treatment interaction. In our example, this is because the mean value of the running variable coefficient is larger than the running variable treatment interaction coefficient.

There is a bias variance trade off between the FIRC One and the other models. The FIRC One model estimate has less error than the other models across the different sample sizes we evaluated (Figure 5). Therefore using the FIRC One model may be justified when the number of total sites is below 30, or the average number of in-bandwidth observations is below 25, regardless of whether there is variance in the running variable coefficients. When the sample size is that small, the RMSE is largest, and there is substantial bias in the other models, eliminating the bias variance trade off. However, when the sample size exceeds either 30 total sites or an average number of in bandwidth observations of 25, the FIRC One model bias leads to a coverage rate for the FIRC One model is well below 95% (Figure 6). Therefore, despite having a lower RMSE, the FIRC One model is not a good choice when there is variance in the running variable coefficients and the sample size is sufficiently large.

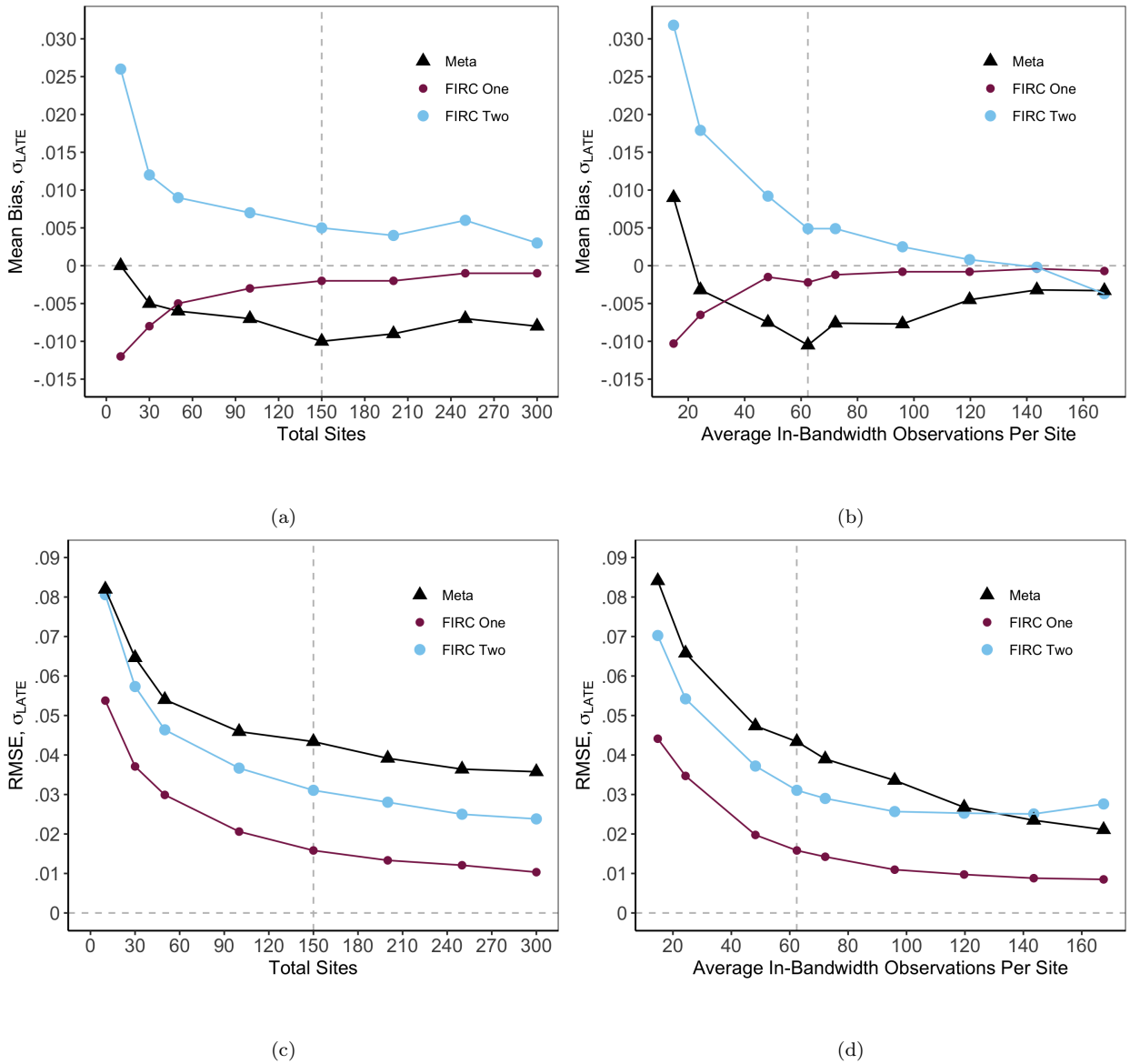
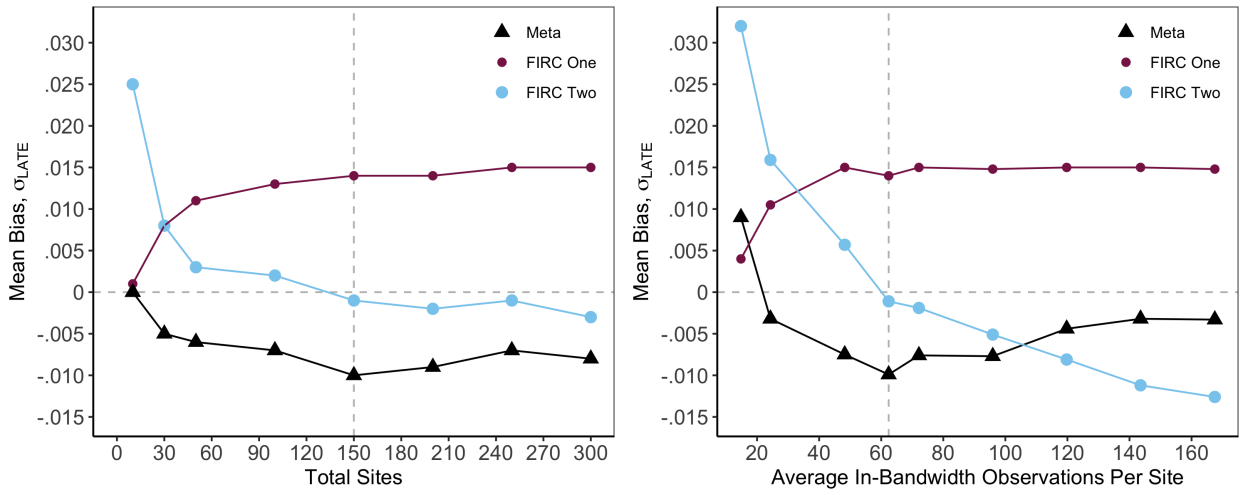
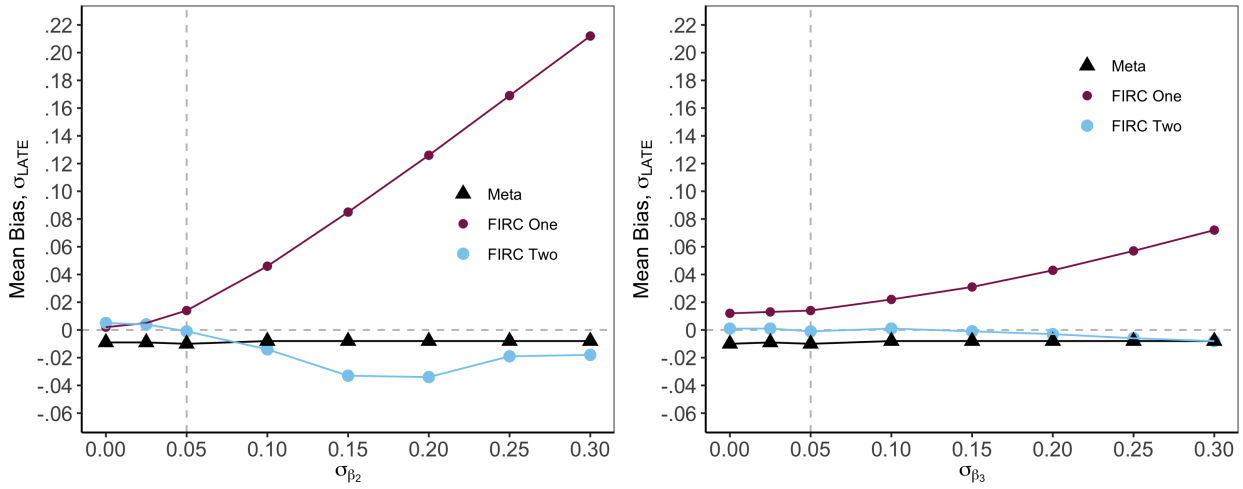


Figure 2: The mean bias (top) and root mean squared error (bottom) in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity models when there is no cross-site variance in the running variable coefficients (i.e., $\sigma_{b0=0}$ and $\sigma_{b1} = 0$). 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 all panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 62 average in-bandwidth observations per site (right).



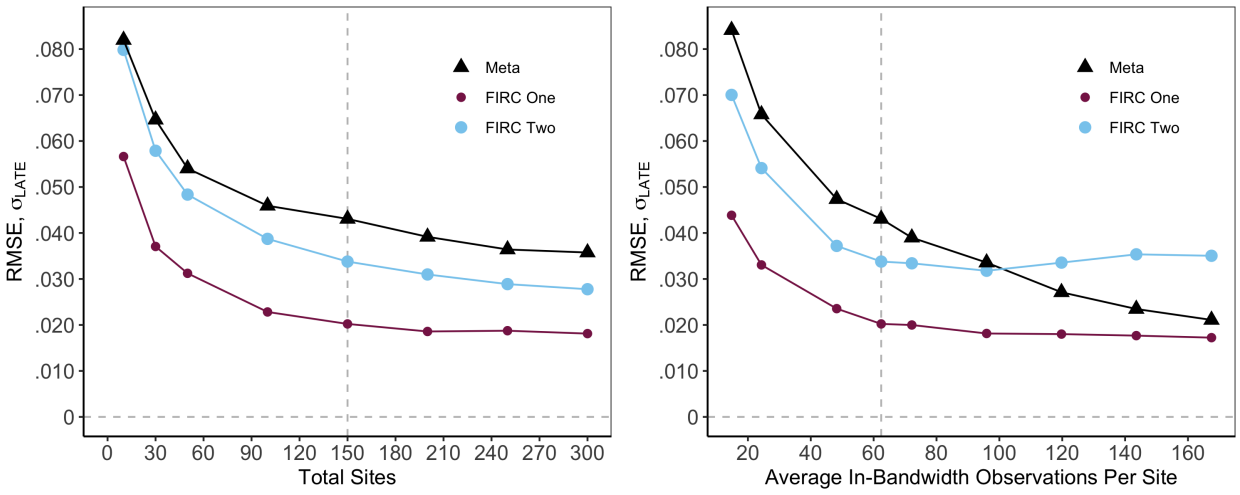
(a) (b)

Figure 3: 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 when there is variance in running variable coefficients ($\sigma_{b0} = .05$ and $\sigma_{b1} = .05$). 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 62 average in-bandwidth observations per site (right).



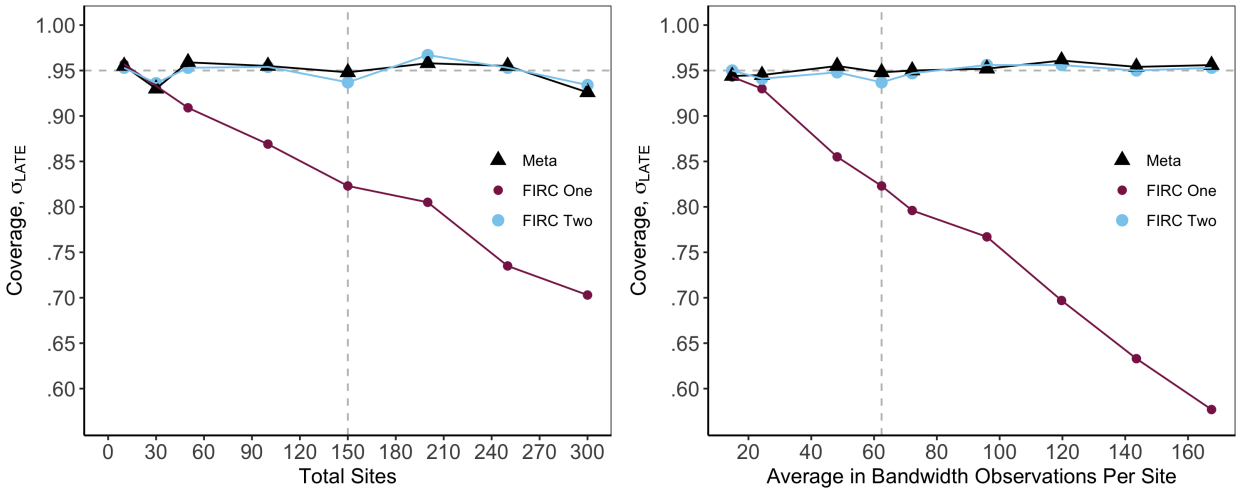
(a) (b)

Figure 4: 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. 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



(a) (b)

Figure 5: The root mean squared error 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 when there is variance in the running variable parameters ($\sigma_{b0} = .05$ and $\sigma_{b1} = .05$). 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 all panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 62 average in bandwidth observations per site (right).



(a) (b)

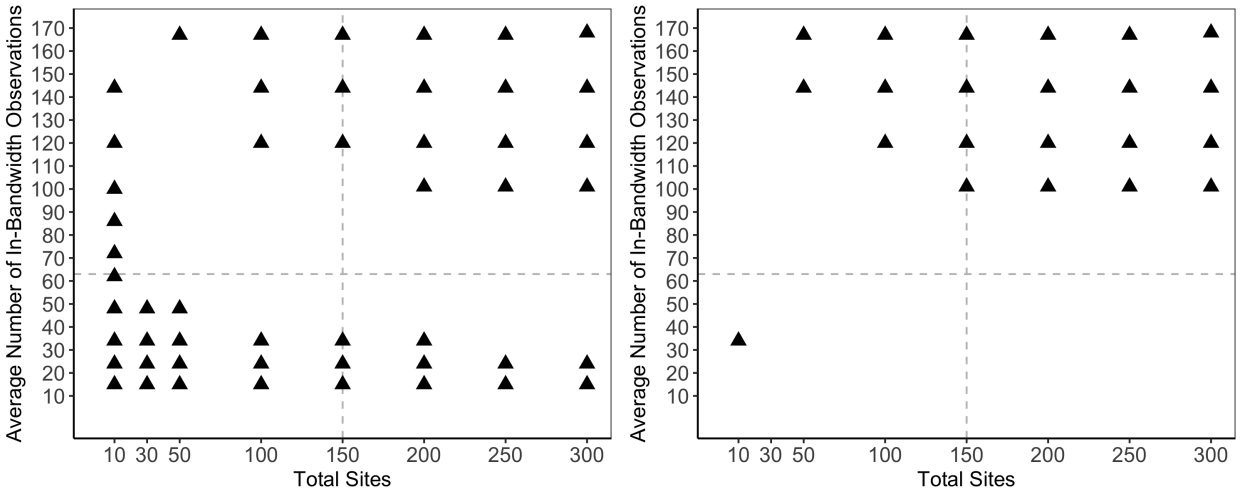
Figure 6: The coverage rate of the cross-site treatment standard deviation confidence intervals for the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity models. In both panels the standard deviations of the running variable parameters (σ_{b0} and σ_{b1}) are set to .05. For all three models, the confidence intervals are estimated using Q-statistic inversion. 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 62 average in-bandwidth observations per site (right).

The Meta Model

If there is cross-site variation in the running variable coefficients, then the Meta model works better than the FIRC Two model when the average number of in-bandwidth observations per site is large. The threshold for determining when the average number of in-bandwidth observations should be considered large depends on the total number of sites in the sample. At the baseline value of 150 total sites, the Meta model starts to consistently have less bias and less error when there are at least 120 in bandwidth observations per site (Figure 3b and 5b).

More generally, we show all the sample size combinations where the Meta model has less bias (Figure 7a) and less error (Figure 7b) than the FIRC Two model. The more total sites in the sample, the lower the large site threshold, where the large site threshold is defined as the point where the Meta model has less bias and error than the FIRC Two model as long as the site size is greater than or equal to that point. When there are 100 or 150 total sites, the large site size threshold is 120 in-bandwidth observations per site; when there are at least 200 total sites, that threshold drops to 100. On the other hand, the large site size threshold gets very large when there are 50 or fewer sites, with 170 being the large site threshold when there are 50 sites. Below 30 sites, the large site threshold is outside our simulation sample, but it also no longer important for model selection because the FIRC One model performs better than either the Meta model or the FIRC Two model, even when there is variance in the running variable coefficients.

The Meta model also has less bias than the FIRC Two model when the site sizes are small, although the FIRC Two model still has less error than the Meta model at these points. However, typically the FIRC One mode performs better than either the Meta model or the FIRC Two model at these small sample sizes regardless of whether there is cross-site variance in running variable coefficients. This makes the small site size threshold less important than the large site size threshold for model selection.



(a) Mean Bias (b) RMSE

Figure 7: Sample sizes where the random effects meta-analysis model regression discontinuity model has less bias (left) or less error (right) than the fixed intercepts random coefficients with random running variable coefficients (FIRC Two) regression discontinuity model, when the cross-site standard deviations of the running variable parameters (σ_{b0} and σ_{b1}) are set to .05.

Summary of Simulation Results

The FIRC One model estimates of the cross-site treatment effect standard deviation have less bias and less error as long as there is no variance in the running variable coefficients. Researchers estimating cross-site treatment effect variance in multisite RDDs, therefore, must test for variance in the running variable coefficients when choosing an estimation model. If cross-site variance in the running variable coefficients is detected, researchers should use the Meta model when the average site size is large, approximately when the site size is above 120 in-bandwidth observations per site, and the FIRC Two model otherwise.

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.

Education Proficiency Plan Evaluation 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 8). 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.

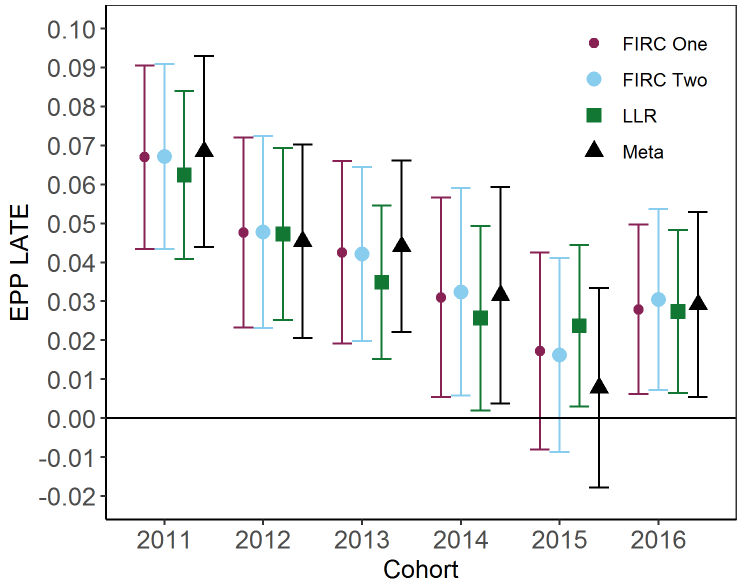
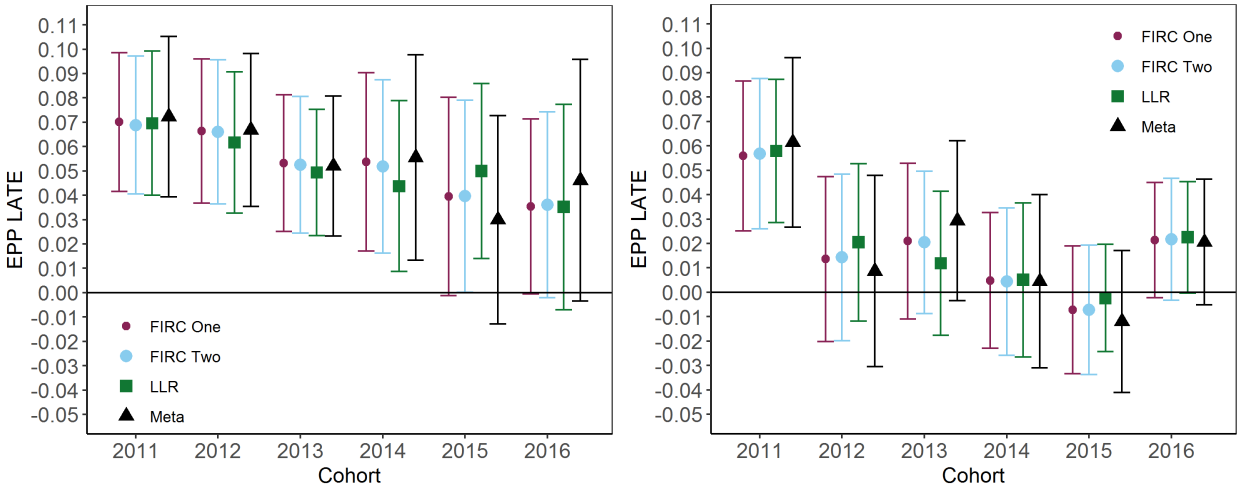


Figure 8: The local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts.

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 9). 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 9: 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 local linear model that is generally used to estimate average treatment effects in multisite RDDs doesn't produce different estimates than the other models.

We use the results from the simulations to select the model for estimating the cross-site treatment effect standard deviation in each cohort. We first fit both FIRC models. In the cases where the FIRC One model had a lower AIC, we used the FIRC One model. In cases where the FIRC Two model had the lower AIC, we would have used the Meta model if the average number of in-bandwidth observations per site was above the large site threshold, but the average number of in-bandwidth observations per site was well below 100 in all of our analyses, and therefore we used the FIRC Two model in all models where we detected variance in the running variables. In all cases, we present results for all three models and mark the estimate from our preferred model in red.

While the math EPP's average effect fell across cohorts, there is persistent cross-high school treatment effect variation across cohorts. Across the six cohorts, the cross-site treatment effect standard deviation is between 7 and 9 percentage points, even as the local average treatment effect is dropping (Figure 10) and is statistically significant in all six cohorts. If we assume that the treatment effect is normally distributed across schools, then these cross-high school treatment effect standard deviations imply that even in the 2011 cohort, when the treatment effect was largest, in more than a third of 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. On the other hand, even as the local average treatment effect of the being required to complete an EPP was dropping, it still had a positive effect on senior year math course completion in many Massachusetts high schools. 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.

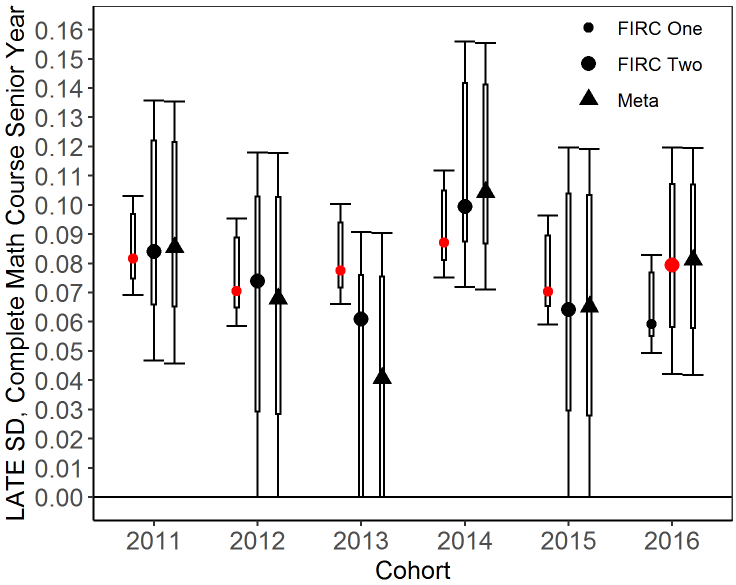
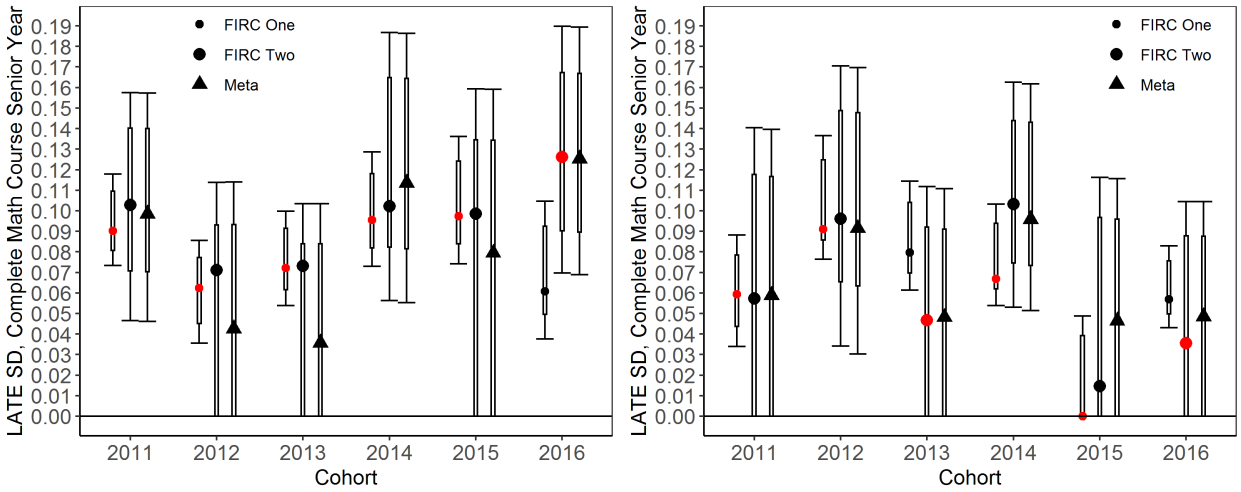


Figure 10: The cross-high school standard deviation of the local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts. The 95% and 80% confidence interval is marked for each point. In each cohort, the estimate marked in red is the preferred model.



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

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

Figure 11: 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 95% and 80% confidence interval is marked for each point. In each cohort, the estimate marked in red is the preferred model.

RDD papers often capture treatment variation by looking at treatment effect heterogeneity by observable characteristics, as in our analysis in Figure 9. However, this type of heterogeneity analysis only captures the "systematic component" of treatment effect variation (Ding, Feller, & Miratrix, 2019). Even though we may have a covariate that explains almost all of the treatment effect, there can still be unexplained treatment effect variation left over. As is the case with the EPP example, we know there is still variation unexplained by our systematic analysis because there is cross-site treatment effect variation in both groups of high schools (Figure 11). Therefore there are still differences in how the policy operates across high schools not fully explained by their high school graduation requirements.

There is more cross-site treatment effect variation amongst the high schools that didn't require four years of math to graduate than those that did. For schools that didn't require four years of math to graduate, the cross-site treatment effect standard deviation ranges from 6 percentage points to 13 percentage points and is statistically significant in all cohorts. In schools that did require four years of math to the cross-site treatment effect, standard deviations range from 0 percentage points to 9 percentage points, and the 95% confidence interval contains zero in

three of the cohorts.

It is not surprising that the cross-site treatment effect standard deviation was larger in the high schools that did not require four years of math to graduate. High schools could require students to complete their EPP either by completing a math course their senior year or by passing a new proficiency exam. Therefore, we expect some of the high schools not generally requiring four years of math to graduate to have required their EPP students to take math their senior year, and some of these high schools not to have requires their EPP students to take math their senior year, which creates cross-site variation.

Among the high schools requiring four years of math to graduate, it is unexpected that the cross-site treatment effect standard deviation is statistically significant in three cohorts. There is no straightforward mechanism for this variance. This demonstrates the usefulness of the cross-site treatment effect standard deviation as a diagnostic for determining parts of an intervention or policy that require more investigation.

Across the eighteen models we ran, in all but four, the FIRC One model had a better fit than the FIRC Two model and was our preferred model. In the simulations, we demonstrated that as long as there is no cross-site variance in the running variable coefficients, the FIRC One model has less bias and error than the FIRC Two model. Our empirical example shows that the FIRC One model also consistently has a shorter interval length than the FIRC Two and Meta models. Looking at the four models, we estimated where the FIRC Two was our preferred model, in two of the models, the FIRC One model estimate of the cross-site treatment effect standard deviation was larger, and in the other two models, the FIRC Two model estimate of the cross-site treatment effect standard deviation was larger. While on average, the FIRC One cross-site treatment effect standard deviation estimates are upwardly biased when there is variation in the cross-site running variable coefficients, there was variation in our simulations, and often the FIRC Two cross-site treatment effect standard deviation estimate would be the larger of the two, which is consistent with what we see in our example.

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 multisite studies (Raudenbush et al., 2012; Reardon & Raudenbush, 2013; Reardon, Unlu, Zhu, & Bloom, 2014; Bloom et al., 2017). In this paper, we show that adapting these methods to the RDD setting is complicated, and methods that may work in the context of RCTs do not work the same within RDDs. Using simulation, we show that a restricted FIRC model should be used when there is no cross-site variance in the running variable coefficients. However, when there is variance in the running variable coefficients, the Meta model should be used when the average number of in-bandwidth observations per sites is large and otherwise the unrestricted FIRC model should be used.

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 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.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, *114*(525), 304–317.

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.

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

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.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes

are unequal. *Biometrika*, *58*(3), 545–554.

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.

Reardon, S. F., & Raudenbush, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociological Methods & Research*, *42*(2), 143–163.

Reardon, S. F., Unlu, F., Zhu, P., & Bloom, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 53–86.

Shapiro, A. (2020, July). *Over diagnosed or over looked? the effect of age at time of school entry on students receiving special education services* (No. 259). Retrieved from `http://www.edworkingpapers.com/ai20-259`

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.

Appendix A Evaluation of the Partially Restricted FIRC Model

The partially restricted FIRC model can be written as follows:

FIRC Three (partially restricted)

Level One - Observation:

$$Y_{ij} = \alpha_j + \beta_{1j}T_{ij} + \beta_{2j}(Score_{ij} - Score_c) + \beta_3 T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij}$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_y^2)$$

Level Two - Site:

$$\beta_{1j} = \delta + e_{1j}$$

$$\beta_{2j} = \gamma_2 + e_{2j}$$

$$\begin{pmatrix} e_{1j} \\ e_{2j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\beta_1} & \sigma_{\beta_1\beta_2} \\ \sigma_{\beta_2\beta_1} & \sigma^2_{\beta_2} \end{pmatrix}\right]$$

As with the other FIRC models, δ is the local average treatment effect and the model is fit using REML.

The partially restricted FIRC model (FIRC Three) allows for variance in the coefficient on the running variable but not the coefficient on the treatment running variable interaction coefficient. Therefore, this model is only misspecified if there is variance in the treatment running variable interaction coefficient, and so it may seem like a reasonable compromise between the fully restricted FIRC One and the fully unrestricted FIRC Two model. However, the FIRC Three model also doesn't perform better than the FIRC Two model even when there isn't variation in the treatment running variable interaction coefficient. Figure 12 shows the bias and RMSE for the different model estimates of the cross-site treatment effect standard deviation across a range of sample sizes when we set the cross-site standard deviation of the treatment running variable interaction coefficient to zero. The FIRC Two and FIRC Three estimates perform comparably across the different sample sizes. The FIRC Two model actually has less mean bias across the different number of total sites when the average number of in-bandwidth observations per site is at its benchmark value of 62.

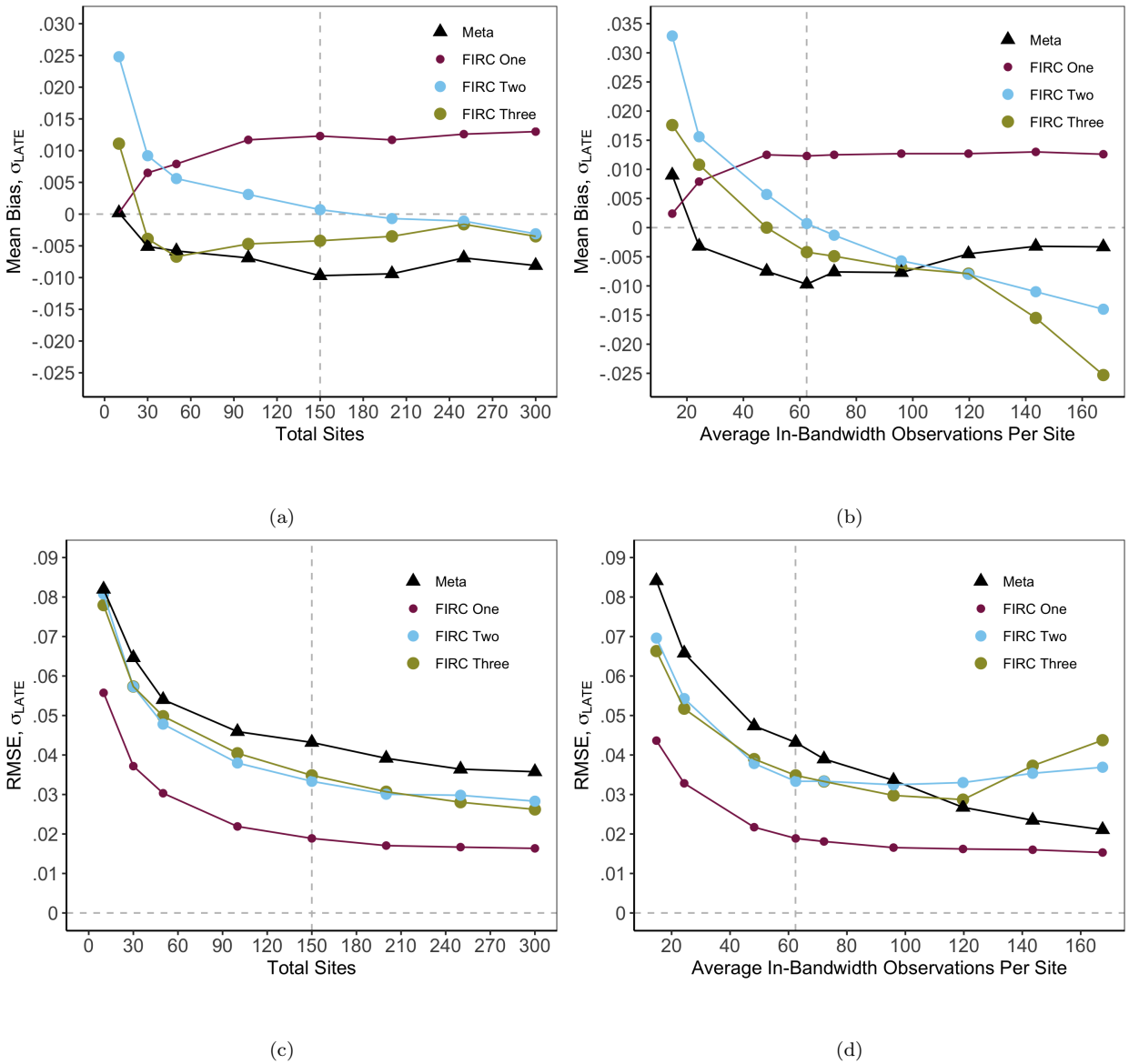
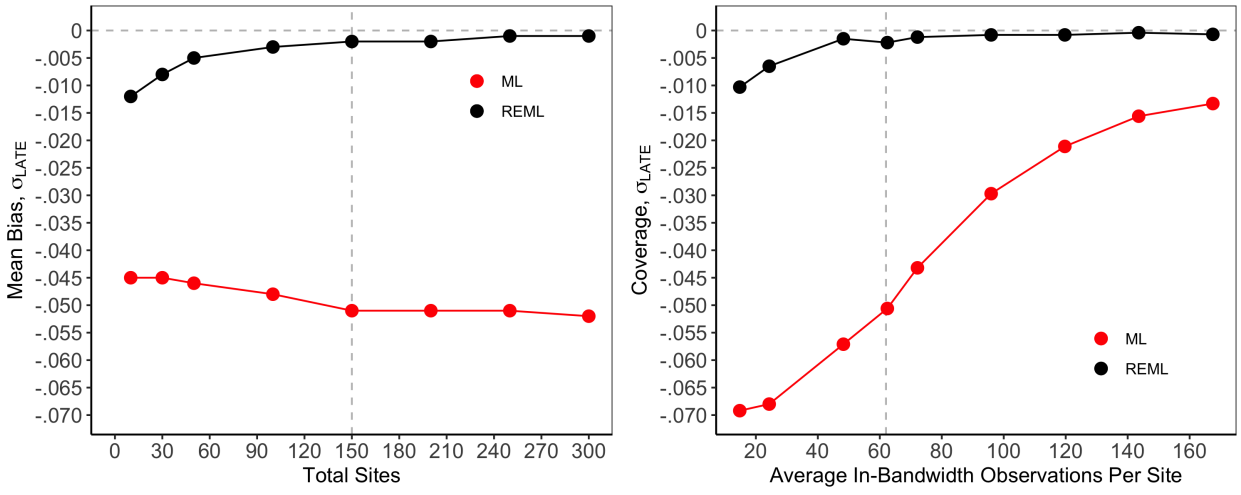


Figure 12: The mean bias (top) and root mean squared error (bottom) in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), fixed intercepts random coefficients with random running variable coefficients (FIRC Two), and fixed intercepts random coefficients with random running variable coefficient and fixed running variable treatment interaction coefficients (FIRC Three) regression discontinuity models when there is cross-site variance in the running variable coefficient and no cross-site variance in the treatment running variable interaction coefficient (i.e., $\sigma_{b0} = .05$ and $\sigma_{b1} = 0$). 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 all panels, the dotted line is at the baseline parameter values of 150 total sites (left) and 62 average in-bandwidth observations per site (right).

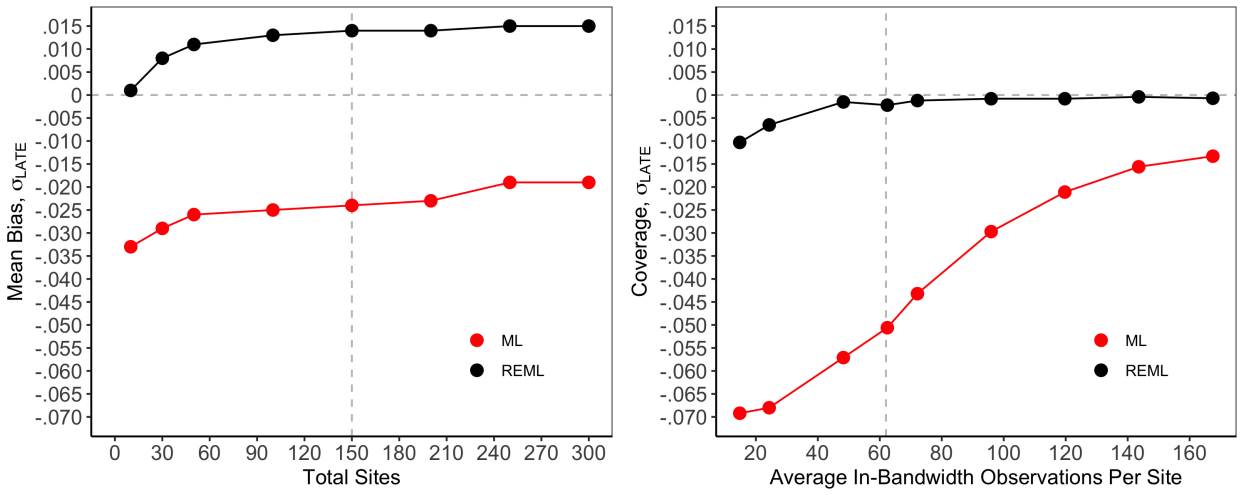
Appendix B Evaluation of Maximum Likelihood Model Estimation

In our main analysis, we use Restricted Maximum Likelihood (REML) to estimate the FIRC models. Maximum Likelihood (ML) estimates of variance parameters in multi-level models are downwardly biased, and REML provides a degrees of freedom correction to remove this bias (Patterson & Thompson, 1971). However, none of the prior multi-site RDD studies that used a multi-level model to estimate cross-site treatment effect variance report whether they used REML or ML to fit their models (Raudenbush et al., 2012; McEachin et al., 2020; Shapiro, 2020). Figure 13 shows that when the restricted FIRC One model is correctly specified, there is a large downward bias to the ML estimates compared to the REML estimates across sample sizes. Similarly, in Figure 14 we show that when the unrestricted FIRC Two is correctly specified, there is also a large downward bias to the ML estimates compared to the REML estimates across sample sizes.



(a) (b)

Figure 13: The mean bias in the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with pooled running variable coefficients (FIRC One) using Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation. In both panels the standard deviations of the running variable parameters (σ_{b0} and σ_{b1}) are set to 0. In the left panels, 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 the both panels the dotted line are at the baseline parameter values of 150 total sites (left) and 62 average in-bandwidth observations per site (right).



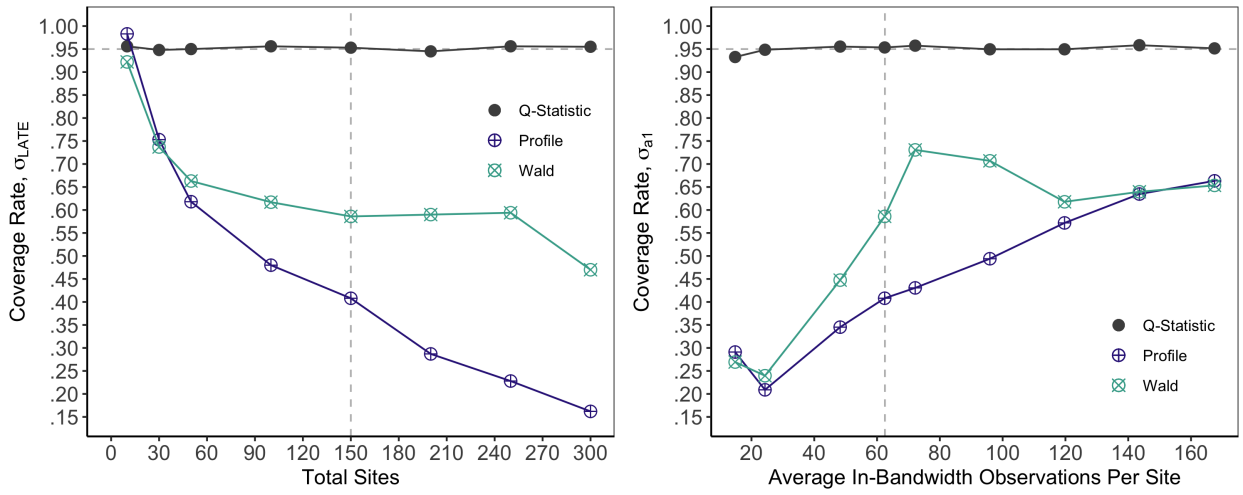
(a) (b)

Figure 14: The mean bias in the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with random running variable coefficients (FIRC Two) using Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation. In both panels the standard deviations of the running variable parameters (σ_{b0} and σ_{b1}) are set to .05. In the left panels, 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 62 average in-bandwidth observations per site (right).

Appendix C Evaluation of Different Confidence Interval Methods for the FIRC Models

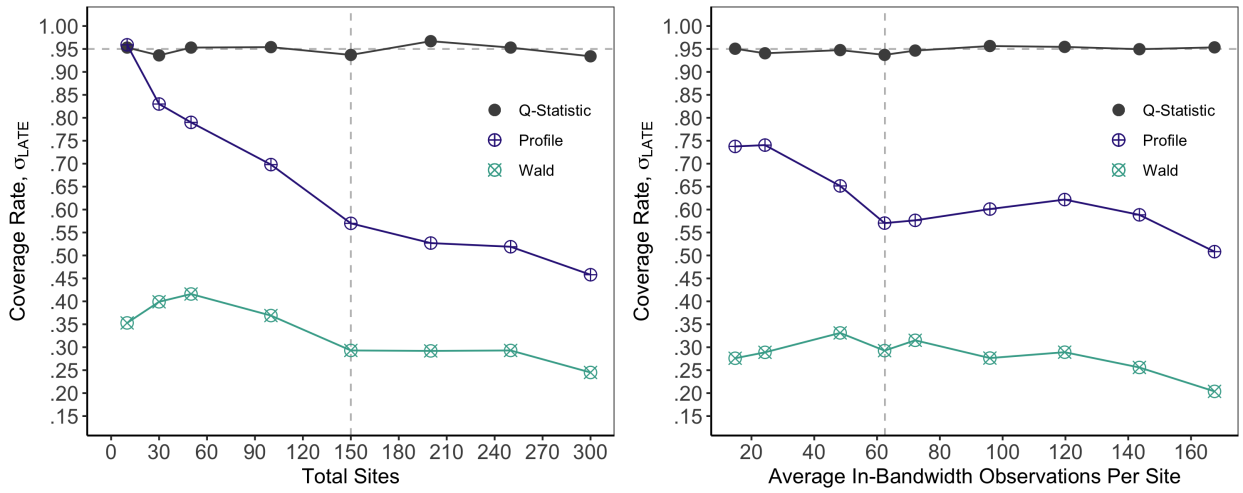
We tested three methods for estimating a confidence interval for the RDD FIRC cross-site treatment effect standard deviations estimates: Wald, Profiled, and Q-Statistic Inversion. In Figure 15 and Figure 16 we show the coverage rate for all three confidence interval types for the FIRC One and FIRC Two models when each model is correctly specified.

Across sample sizes, the coverage rate for both the Wald and profiled confidence intervals is well below 95%. Wald confidence intervals are known not to work well for variance components of multi-level models, and so it is not surprising that they work poorly in the RDD FIRC model. The issue with the profiled confidence intervals is more complicated. Profiled confidence intervals are obtained by performing test inversion on the likelihood ratio test; the likelihood function is estimated for a range of cross-site treatment effect standard deviation values and compared to the maximum likelihood cross-site treatment effect standard deviation estimate. The 95% confidence interval is all values where we cannot reject the null hypothesis that the fits of the two values are equally good. Profiled confidence intervals only work with maximum likelihood estimates and not degrees of freedom corrected REML estimates because the profile method compares values of the likelihood function. This creates a problem because the non-degrees of freedom corrected maximum likelihood estimates have a large downward bias (See Appendix B), and therefore the confidence intervals generated around this biased estimate have poor coverage. The Q-statistic inversion method consistently produces 95% confidence intervals with a coverage rate of 95%, and so this is the method we use in our analysis.



(a) (b)

Figure 15: The coverage rate of the Wald, Profiled, and Q-Statistic Inversion confidence intervals of the cross-site treatment effect standard deviation estimated from the fixed intercepts random coefficients with pooled running variable coefficients (FIRC One) model. In both panels the standard deviations of the running variable parameters (σ_{b0} and σ_{b1}) are set to 0. 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 62 average in-bandwidth observations per site (right).

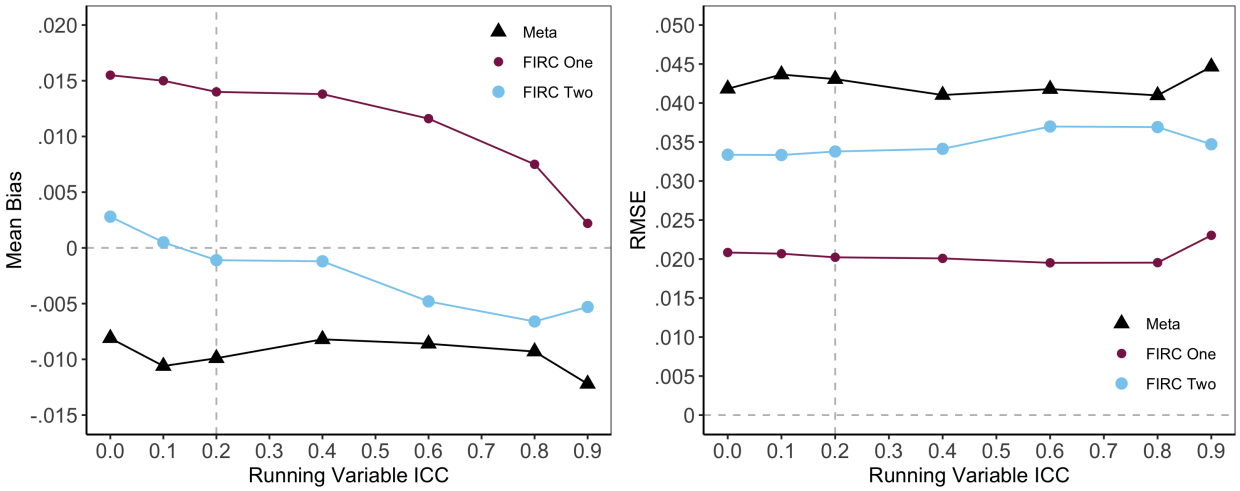


(a) (b)

Figure 16: The coverage rate of the Wald, Profiled, and Q-Statistic Inversion confidence intervals of the cross-site treatment effect standard deviation estimated from the fixed intercepts random coefficients with random running variable coefficients (FIRC Two) model. In both panels the standard deviations of the running variable parameters (σ_{b0} and σ_{b1}) are set to .05. 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 62 average in-bandwidth observations per site (right).

Appendix D Simulation Results by Interclass Correlation

We find that changing the interclass correlation of the running variable doesn't substantively change our results. For all three models, the mean bias and root mean squared error stays approximately constant as the running variable interclass correlation is varied from 0 to .9 (Figure 17). The only major exception is that when the running variable interclass correlation is .9, the mean bias of the restricted FIRC model (FIRC One) gets close to zero.



(a) (b)

Figure 17: The mean bias (left) and root mean squared error (right) 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 by the running variable interclass correlation. The cross-site variance in the running variable coefficient and cross-site variance in the treatment running variable interaction coefficient (σ_{b0} and σ_{b1}) are fixed at .05 , the average number of observations per site is fixed at 130 (62 average in-bandwidth observations), and the number of total sites is fixed at 150.