Three Essays on Making Casual Inferences with Test Scores

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Education

have examined a dissertation entitled

Three Essays on Making Casual Inferences with Test Scores

presented by Sophie Lilit Litschwartz

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Andrew Ho

Signature _____

Typed name: Prof. Luke Miratrix

Signature _____

Typed name: Prof. Eric Taylor

Signature _____

Typed name: Prof.

Signature _____

Typed name: Prof.

Date: May 10, 2021

Three Essays on Making Casual Inferences with Test Scores

A dissertation presented

by

Sophie Lilit Litschwartz

to

Committee on Higher Degrees in Education

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Education

Harvard University

Cambridge, Massachusetts

May 2021

Dissertation Advisors: Andrew Ho, Luke Miratrix Sophie Lilit Litschwartz

Three Essays on Making Casual Inferences with Test Scores

Abstract

In education research test scores are a common object of analysis. Across studies test scores can be an important outcome, a highly predictive covariate, or a means of assigning treatment. However, test scores are a measure of an underlying proficiency we can't observe directly and so contain error. This measurement error has implications for how we use test scores in research. In this dissertation, I combine psychometrics and causal inference to develop three methods for doing education research with test scores.

In the first study, I combine Classical Test Theory and simulation to develop a generalized method for adjusting test score distribution where there was a policy to either selectively retest or rescore initially failing students. Using this method, I show how adjusting for retesting on a North Carolina accountability exam reduces the estimate of mean growth across testing occasions from .17 standard deviations to near zero. I also reexamine an investigation of "score scrubbing" on the New York Regent and demonstrate rescoring can inflate perceived scrubbing rates by a factor of three, from 12% to 36%.

The second and third studies contribute the literature on regression discontinuity design. In the second study, I create and evaluate two methods for estimating cross-site treatment effect variation in multi-site RDDs, one based on random-effects meta analysis and the other based on the fixed intercepts random coefficients model. I use these models to evaluate Massachusetts's "Education Proficiency Plan" policy and find enough treatment effect variance in three cohorts for the treatment effect to have been negative in more than a third of high schools. In the third study, I apply a psychometric latent variable framework to regression discontinuity design and derive the amount biased induced by analyzing a

regression discontinuity design using a local randomization framework.

Contents

List of Tables

List of Figures

Acknowledgments

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA-NSF Grants Program" from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those AERA or NSF. This research was also supported in part by the Institute of Educational Sciences, U.S. Department of Education, through grant R305B150010 to Harvard University. The opinions expressed are those of the author and do not represent the views of the Institute for Education Sciences or the U.S. Department of Education.

If I have learned anything studying education, it is that individual educational achievements require the support of a great many people, and this dissertation is no exception. I would like to thank my dissertation committee, Andrew Ho, Luke Miratrix, and Eric Taylor, for teaching me what makes a good research question, providing feedback on multiple paper drafts, and guidance throughout my doctoral career. I would also like to thank Felipe Barrera-Osorio, Carrie Conaway, the C.A.R.E.S. lab, the measurement lab, and the HGSE colloquium for providing helpful comments on my work. A special thanks to Nicole Pashley, who gave me the idea for my third paper.

I am thankful to Elana McDermott, Matt Deninger, Adrienne Murphy, Andrew Martin, Bob Lee, Nyal Fuentes, and Kathryn Sandel at the MA Department of Elementary and Secondary Education for contextualizing the Education Proficiency Plan policy for me and providing data. I would also like to thank the Center for Education Policy Research for facilitating data access and storage. I am also grateful to Irene Pak and Ashley Dixon at the PIER fellowship for all their support.

I would not have ended up in graduate school in the first place without many great teachers throughout my pre-graduate education. I would especially like to thank my middle school math teachers, Mr. Delepine and Mr. Garcia, for teaching me to push myself on challenging material and Randall Reback for introducing me to education policy research and rigorously thinking through policy questions.

I have been lucky to have the support of many friends and colleagues. I am grateful for the many hours spent tackling problem sets, working through ideas on the office whiteboard, the evenings spent in the trampoline park, the lunches in the Gutman Cafe, the supportive conversations, and the clear advice. I'm especially lucky to have non-grad school friends who were always willing to let me talk about my research or reassure me through my anxieties.

This dissertation would not have been possible without Flavia Silva, and Morris and Elijah's teachers at Pooh and Friends, who watched and nurtured my children under unprecedented circumstances. I would not have had the time to write this without them.

Finally, I would like to thank Ari for his support, love, and insight. I could not have asked for a better partner.

For Morris and Elijah, without you I would not be the person who wrote this

Introduction

In education research, test scores are a common object of analysis. Across studies, test scores can be an important outcome, a highly predictive covariate, or a means of assigning treatment. However, test scores are a measure of underlying proficiency that can not be observed directly and therefore measure this proficiency imperfectly and with error. Furthermore, details about testing policies and score construction can radically shift the distribution of this measurement error. The unique statistical properties of test scores have implications for analysis. Therefore, causal inference in education requires statistical methods specifically built with the statistical properties of test scores in mind.

Developing causal inference methods for use with test scores benefits from an interdisciplinary perspective. Psychometrics, a field concerned with the quantification of mental attributes and the statistical properties of test scores, is a branch of psychology (American Psychological Association, 2021). Causal inference is a field concerned with understanding how changing conditions (e.g., a new policy or treatment) change outcomes (Pearl, 2009) and spans academic disciplines. Creating statistical methods for doing causal inference research with test scores, therefore, requires pulling statistical tools from across disciplines together and forming them into methods designed for performing causal inference with test scores. With this in mind, I present three essays that pull from across psychometrics, statistics, econometrics, and meta-analysis to create new methods for answering causal research questions in education that involve test scores.

In the first essay, I develop a method for adjusting test score distributions from tests where there was a policy to either selectively retest or rescore initially failing students. In

pass/fail exams, policymakers care about minimizing the number of test takers who meet the proficiency standard but fail the exam due to measurement error (false failures), and test takers who meet the proficiency standard but pass the exam due to measurement error (false passers). Policymakers are often more concerned with false failures and therefore will selectively rescore or retest when a test taker has failed. However, selectively retesting or rescoring distorts the test score distribution and changes the interpretation of the mean, variance, pass rate, and even the histogram.

I combine Classical Test Theory and simulation to allow for the adjustment of any distribution level quantity to account for retesting and rescoring. I apply this method to two examples: a 3rd-8th grade accountability exam in North Carolina that required initially failing test takers to retest two weeks later and an investigation of “score scrubbing" on a high school exit exam in New York that rescored all initially failing exams near the cut point. I show how adjusting for retesting on the North Carolina accountability exam reduces the estimate of mean growth across testing occasions from .17 standard deviations to near zero and demonstrate how rescoring in New York was able to inflate perceived scrubbing rates a factor of three, from 12% to 36%.

The second essay, with Luke Miratrix, formalizes and evaluates methods for estimating cross-site treatment effect variance in a multisite regression discontinuity design. Quantifying treatment effect variation is 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. In multisite studies, cross-site treatment effect variance is one way to quantify treatment effect variation. While there are standard methods for estimating the cross-site treatment effect variance in experiments, there are no such methods for regression discontinuity designs.

The first part of this essay develops two methods for estimating cross-site treatment effect variance in multisite regression discontinuity designs: one that combines the fixed intercept random coefficient (FIRC) model with the regression discontinuity local linear regression model and another based on random effects meta-analysis (Meta). Using simulation, we

show that a restricted FIRC model works best as long as there is no cross-site variance in the running variable coefficients. When there is cross-site variance in the running variable coefficients, an unrestricted FIRC model works best when the average number of in bandwidth observations per site is less than 100, and the Meta model works best when the average number of in bandwidth observations per site is more than 100.

In the second part of this essay, we apply the models for estimating cross-site treatment effect variance to an education policy that used test scores as a running variable. In Massachusetts, students who pass the high school exit exam but are still determined to be nonproficient must 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 consistently large enough for the treatment effect to have been negative in more than a third of high schools. We also find that, while differences across high schools in graduation requirements were able to explain all of the treatment effect, they were not sufficient to explain all of the treatment effect variance.

In the third essay, I apply a classical test theory model to local randomization regression discontinuity analysis. Explanations of the internal validity of regression discontinuity design studies generally appeal to the idea that regression discontinuity designs are “as good as" random near the treatment cut point. Cattaneo, Frandsen, and Titiunik (2015) are the first to take this justification to its full conclusion and propose estimating the regression discontinuity design local average treatment effect the same way one would a randomized experiment. I this essay, I explore what analyzing a regression discontinuity design as a local random experiment would mean when the running variable is a test score. I derive a formula for the bias in the local average treatment effect estimate estimated using the local randomization method, $a\rho\Delta$. Where a is the relationship between latent proficiency and the potential outcome absent treatment, ρ is the test reliability, and Δ is the distance between the treatment and control running variable value. I argue that this bias will make local randomization problematic for most regression discontinuities that use a test score as

the running variable.

Collectively, these three essays enlarge the statistical methods toolbox for education research and contribute to a long tradition of interdisciplinary methods.

Chapter 1

A General Method for Adjusting Test Score Distributions to Account for Rescoring and Retesting

Abstract

Exams frequently have testing policies that rescore or retest only initially failing examinees. This asymmetric treatment of passing and failing distorts the test score distribution and can bias analysis of test scores. In this study, I develop a method that combines Classical Test Theory and simulation which removes the bias induced by rescoring and retesting from any distribution level quantity. Using this method, I show how adjusting for retesting on a North Carolina accountability exam reduces the estimate of mean growth across testing occasions from .17σ to near zero. I also reexamine an investigation of “score scrubbing" on the New York Regent and demonstrate rescoring can inflate perceived scrubbing rates by a factor of three, from 12% to 36%.

1.1 Introduction

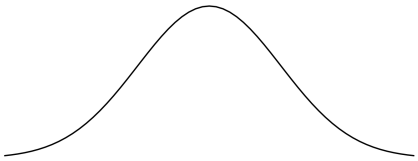
It is a common practice for standardized exams to have a process for rescoring exams and retesting examinees. When exams are used to make a pass/fail determination, exam programs will frequently only rescore or retest initially failing exams/examinees. To give just a few examples: the GED exam, all United States high school exit exams, all United States legal bar exams, the United States Medical Licensing Exam, the National Council Licensure Examination for nurses in the United States, and United States Naturalization Exam all allow initially failing students at least one opportunity to retest. In the important case of high school exit exams, retesting is explicitly recommended by The Standards for Educational and Psychological Testing (AERA, APA, and NCME, 2014).

Selective rescoring is also commonly done on pass/fail exams with uncertainty in the grading process. In a comprehensive review of rescoring policies for the thirteen states (Florida, Indiana, Lousiana, Massachusetts, Maryland, Mississippi, New Jersey, New Mexico, New York, Ohio, Texas, Virginia, Washington) that required high school exit exams for the class of 2020, I found that all but four (Massachusetts and the PARCC states Maryland, New Jersey, and New Mexico) had a process that allows for the rescoring of some exams that failed to achieve a passing score. Rescoring initially failing examinees is also a standard practice of state bar admittance exams. A 2011 survey found that 23 state legal bar associations automatically rescore either all failing bar exams or failing bar exams near the minimum passing score (Albanese, 2016).

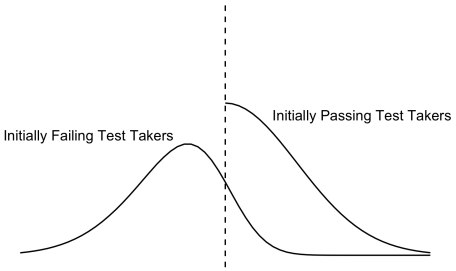
Rescoring and retesting distorts the test score distribution, creating opportunities for confusion when researchers and the public do secondary analysis on these distributions. In one specific case, reporting by the Wall Street Journal in 2011 showed large discontinuities in the New York City Regent test score distribution around the passing cutoff (Martinez and McGinty, 2011). The Wall Street Journal claimed, and academic researchers later concurred, that the discontinuity was proof that teachers were purposely manipulating or “scrubbing" test scores near the cutoff (Dee, Dobbie, Jacob, and Rockoff, 2019). However, prior to this news article, it was official policy to regrade all initially failing exams within five scaled

points of passing. This selective rescoring policy was sufficient to create a discontinuity even absent any “scrubbing".

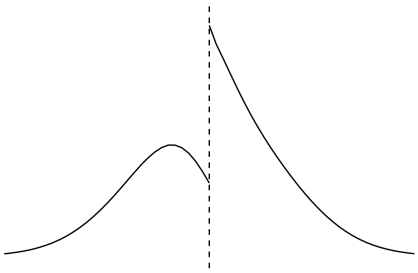
Under a Classical Test Theory (CTT) model, test scores are an additive combination of a student’s true proficiency and measurement error (Lord and Novick, 1968). Rescoring the exam effectively redraws the exam measurement error introduced by the grading process. Only regrading failing exams creates a new distribution of scores for the initially failing examinees while leaving the distribution of the initially passing examinees untouched. The new error draw will cause some of the initially failing students to now have passing scores, but no initially passing students will end up with failing scores. When the new distribution for the initially failing students is combined with the old distribution of initially passing students, there is an excess of mass to the right of the cut point and missing mass to the left of the cut point. Therefore, a policy of selective rescoring failing students creates a discontinuity at the pass point on its own with no score scrubbing (Figure 1.1).



(a) *A smooth test score distribution before any exams are rescored.*



(b) *The scoring error is redrawn for initially failing exams, while the distribution for initially passing exams remains unchanged. The distribution on the left is the new scores for the initially failing examinees with the redrawn error components. The distribution on the right is the unchanged scores for the initially passing examinees.*



(c) *The initially failing and initially passing distributions are combined. This creates a large discontinuity at the exam cut point.*

Figure 1.1: *A stylized model of selective rescoring assuming a test reliability of .9*

A selective rescoring policy leads to a discontinuous distribution because it treats initially failing and initially passing students asymmetrically. Selective retesting also treats initially passing and initially failing test takers differently, which has a similar effect on distributions. Retesting failing test takers, like rescoring failing tests, redraws the measurement error for failing examinees while leaving the measurement error for passing examinees unchanged.

Treating initially failing test takers and initially passing test takers differently is not

inherently wrong. The costs of false failure (i.e., examinees with true proficiencies greater than the pass cut failing due to random error) might be higher than the cost of false passing (i.e., examines with true proficiencies lower than the pass cut passing due to random error). Certification and licensing exams specifically are often necessary and not sufficient conditions for receiving a specific certificate or license. That is, in cases like high school exit exams or legal bar exams, examinees must complete coursework in addition to passing the exam; this may leave policymakers less concerned about false passing than false failing, since passing test takers must also demonstrate proficiency another way but failing test takers automatically don't receive their certification or license. Finally, in the case of retesting, treating failing and passing differently serves an additional purpose: it also gives non-proficient test takers a chance to learn the exam content and become proficient. Therefore, policymakers may want to allow retesting even when they do not treat the costs of false failure as larger than the costs of false passing.

While selective rescoring and retesting policies may be justified, the New York Regent Exam's case shows how exams with rescoring or retesting policies can have distributions that are easily misunderstood. Selectively rescoring or retesting failing students distorts the distribution so that it is hard to disentangle statistical artifact from human behavior. Selective rescoring and retesting also change how researchers need to interpret fundamental distributional descriptive quantities such as the exam mean, standard deviation, pass rate, test standard, and even the histogram.

It is not news to psychometrics that rescoring and retesting must be accounted for when interpreting test scores. However, the existing literature on rescoring and retesting focuses primarily on test design and not on how to do data analysis with test scores from exams with rescoring or retesting policies. Even the literature on test design when there is rescoring or retesting is limited. Articles have been written about the need to adjust for retesting when setting test standards, either in the initial standard setting process or later by test administrators (Geisinger and McCormick, 2010; Millman, 1989). However, these articles have not quantified how much the test standard would need to be adjusted to

account for retesting. Several studies have expressly provided methods for adjusting type misclassification rates (i.e., the rate of students wrongly classified as either passing or failing given their true score) when designing exams with either rescoring or retesting (Bradlow and Wainer, 1998; Clauser and Nungester, 2001; Douglas and Mislevy, 2010; Huynh, 1990). Finally, Cheng and Liu (2016) derive a closed form solution for estimating the mechanical effect of retesting on the pass rates specifically.

In this paper, I develop a method for analyzing test scores when there has been selective rescoring or retesting that combines Classical Test Theory and simulation. This method fills two major gaps in the literature on rescoring and retesting. First, this is a method focused on how to do data analysis on tests after there has been rescoring and retesting. Second, this method allows researchers and policymakers to estimate the direct effect of rescoring and retesting on any distribution level quantity of interest and not just a few special cases. Specifically, I will demonstrate how this method can be used to 1)visualize the discontinuous distributions that result from rescoring and retesting; 2) isolate the direct effect rescoring and retesting have on any distribution level quantity of interest; 3) generate prediction intervals to isolate human behavior from measurement error.

I demonstrate this method by applying it to two exams where rescoring and retesting complicated test score analysis. I use real data examples to make it easier for practitioners to apply this work to their own research problems. These examples are designed to illustrate the versatility of this paper's rescoring and retesting correction method. Each example has different data limitations (e.g., the first example is missing scores outside the retested region and the second example only has the final test scores); in each instance, I estimate the rescoring and retesting effect on different quantities of interest. In both examples, I demonstrate the importance of adjusting for rescoring and retesting to interpret the test scores correctly.

In the first example, I look at an exam retesting policy in North Carolina, where students who initially failed the state accountability exams were required to retest a few weeks later. A recent regression discontinuity study found that students who narrowly failed the exams

and thus were required to retake the state math exam performed .03σ better on the state math exam one year later than the students not required to retake the state exam (Aucejo, Romano, and Taylor, 2020, "ART"). However, Aucejo, Romano, and Taylor (2020) do not measure what happened to test scores between testing occasions, which can provide insight into why the retested students had higher test scores a year later.

I show that while the unadjusted mean test score growth for students at the margin of passing was .17σ between testing occasions, when the statistical effect of retesting is accounted for the mean test score growth drops to near zero. This result, combined with prior research on test score fade out, suggests that differences between retested and non-retested students a year later were not just the result of proficiency gains made by the retested students between testing occasions but also proficiency losses in the non-retested group. I also use this analysis to show that the mean test score growth was the largest for the lowest scoring students, and decreased as initial student test scores increased.

The second example is the New York Regent exam. I estimate the amount of missing mass the rescoring rule alone would have created. I use this to show that some score scrubbing took place on the New York Regent exam, but that naive estimates inflate the amount by a factor of three. I also evaluate a selective rescoring policy's merits by showing how rescoring effects the false fail rate, the false pass rate, and the overall pass rate. Using these metrics, I conclude that the selective rescoring rule was statistically equivalent to lowering the passing threshold by one point.

1.2 Retesting in North Carolina Accountability Exams

In this paper, I show how to combine Classical Test Theory (CTT) and simulation to analyze distributions from tests with selective rescoring and retesting policies. I begin in this section by demonstrating this method in the context of retesting, where the statistical model is more straightforward. The following section shows how this model can be adapted to the more complicated rescoring case.

1.2.1 Analytical Retesting Model

Following CCT, the observed score for examinee i at testing occasion t can be modeled as:

$$X_{it} = \theta_i + \epsilon_{it},\ \epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_{it}}) \tag{1.1}$$

where θ_i is examinee i's true proficiency and ϵ_{it} is the measurement error for examinee i at testing occasion t. This testing measurement error is all test occasion based random error, which includes error associated with the test items sampled (e.g., are the items particularly easy or difficult), testing conditions (e.g., are the distractions in the test room), and random fluctuations in the state of the test taker (e.g., did the student get a good night sleep the night before). This error doesn't account for non-classical forms of error, such as error from inappropriate test preparation activities (Koretz, 2008)

At each testing occasion, the measurement error (ϵ_{it}) is redrawn. The true proficiency (θ_i) is fixed across testing occasions by assumption. Fixing the true proficiency allows the model to test the hypothesis of that there was no change in true proficiency across testing occasions.

Selective retesting distorts the distribution by redrawing ϵ for some students and not for others. This process can be modeled directly using simulation and requires four inputs: the error variance (σ^2_ϵ), a θ distribution, a rule for determining when an examinee is retested, and a rule for converting the vector of an examinees scores at each occasion (X_{it}) into a final score (X_{ifinal}). I refer in this paper to the effect of retesting as the difference between the modeled X_{final} distribution and the initial distribution X_1.

1.2.2 North Carolina Accountability Exams Background

Between 2009 and 2012, North Carolina enforced a policy where students who initially failed a state accountability exam were retested two to three weeks later. Like most states, North Carolina students in 3rd to 8th grade were given ELA and math tests at the end of the year. Based on these tests, students were assigned a proficiency Level between I and IV, where proficiency Levels I and II were considered failing, and Levels III and IV were

deemed to be passing. Under the retesting policy, all students initially categorized as Level II were required to retake the accountability exams to prevent students from failing due to measurement error. For students who retested, the final accountability determinations used the highest of the two scores. In the end, schools were assessed on two metrics: 1) final percent passing and 2) between year test score growth (Aucejo et al., 2020; North Carolina Department of Public Instruction, 2009).

ART looked at the consequences of this retesting policy and found that initially failing the math accountability exam and therefore being required to retake the exam increased a student's math score on the next year's exam by $.03\sigma$. They examine several mechanisms for why being in the retested group caused an increase in future test scores and conclude the most likely explanation is that teachers prioritized teaching the initially failing students over the initially passing students in the weeks between the exams. However, this analysis lacked a method for estimating how much students learned in the two to three weeks between testing occasions; and if this theory about the retesting is correct we would expect to see growth by the retest. In this section, I extend the work of ART and provide a method for measuring student learning between testing occasions.

Drawing on summary statistics of the initial and retested scores for Level II students presented by ART (Appendix Table C2), I model the retesting process. These summary statistics contain the initial empirical distribution for the 389,374 Level II students with recorded scores for the initial test, the retest, and the subsequent year's test. The summary statistics do not contain information for students outside the Level II score range or for 8th graders who were not tested the next year.

The modeling in this example is primarily based on the initial Level II empirical distribution, but the model also requires an estimate of the full observed score mean and variance. A total of 2,083,814 North Carolina students in grades 3-7 took the end of year math accountability exam between 2009 and 2012 and had a recorded score in the following year (E. Taylor, personal communication, July 20, 2020). The summary statistics are a truncated interval of the full distribution of students who would retest if random test error

placed them in the retesting interval. The state required Level II students to retake the exam, and compliance was high: 98% of students retook the exam. The full distribution, therefore, has a total of 2,001,295 observations. I estimate the mean and standard deviation of the full observed score distribution with a grid search. I take all combinations of a mean in [4,10] and a standard deviation in [4,12] in .01 increments. I use the mean and standard deviation combination with the least squared error between density from the pdf at this mean and standard deviation combination and the empirical Level II density at each scaled score point. This grid search produces an estimated observed score variance of 95 and mean of 6.55, as measured in scaled score points centered around the pass/fail cut point.

1.2.3 North Carolina Retesting Model

The retesting model takes four inputs: an estimate of the error variance (σ^2_ϵ), an estimate of the θ distribution, a rule for determining when an examinee is retested, and a rule for converting the vector of an examinees scores at each occasion (X_{it}) into a final score (X_{ifinal}). I next discuss obtaining these values.

1. Estimate of error variance $\sigma^2_{\epsilon_{it}}$:

In this example, I simplify this estimation problem by assuming homoskedasticity across students and occasion. Therefore, I drop the i and t indices, estimating a single σ^2_ϵ. In the CCT framework, the test reliability (ρ) can be defined as $\rho = \frac{\sigma^2_\theta}{\sigma^2_\theta + \sigma^2_\epsilon}$,where σ^2_θ is the true score variance and σ^2_ϵ is the error variance. From this we can estimate the error variance as:

$$\sigma^2_\epsilon = (1 - \rho)\sigma^2_X \tag{1.2}$$

As stated above, I estimate the initial observed score variance (σ^2_X) to be 95. The North Carolina Department of Public Instruction does not currently make exam technical reports available before the 2012-2013 school year, one year after the relevant years in our sample. In the 2012-2013 school year, all math exam reliabilities were between .9 and .94 across forms and grades. The overall average reliability was .92, which I use in the main analysis (Mbella, Zhu, Karkeend, and Lung, 2016). I also run my analysis for all reliabilities between .8 and

.99 in .01 increments to account for both the limited information on test reliability in the sample years and overall uncertainty in reliability estimates.

2. Estimate of the θ distribution:

In this empirical example, I have the initial unretested observed score distribution of the retested students. Following Lord and Novick (1968) and Dudek (1979), in CCT the conditional θ distribution at each observed score value is:

$$\begin{aligned} \theta_i|X_{it} &\overset{iid}{\sim} N(\mu_{\theta_i|X_{it}}, \sigma^2_{\theta_i|X_{it}}) \\ \mu_{\theta_i|X_{it}} &= \rho X_{it} + (1-\rho)\mu_X \\ \sigma^2_{\theta_i|X_{it}} &= \rho(1-\rho)\sigma^2_X \end{aligned} \tag{1.3}$$

The initial observed score mean μ_X is estimated as 6.55. The $\sigma^2_{\theta_i|X_{it}}$ in this equation is sometimes referred to as the standard error of estimation.

I estimate the θ sample from this distribution by drawing a sample of θs for each initial observed score in the retesting region. For example 59,084 students initially score one scaled score point below passing. I model the true score distribution for these students by drawing 59,084 observations from the distribution $N(\rho(-1) + (1-\rho)(6.55), \rho(1-\rho)(95))$. To get the full θ sample, I repeat this process for each scaled score value in the retesting region.

3. A rule for determining when an examinee is retested:

The θ distribution is estimated from an observed score distribution that only includes students who retested. Therefore, I assume that all students with an initial score below the pass/fail cut retest since, by construction, the observed score distribution is the distribution of retesting students. For these retested students, I draw a new distribution of 389,374 error terms ϵ_{i2} and estimate a second observed score as:

$$X_{i2} = \theta_i + \epsilon_{i2} \tag{1.4}$$

4. A rule for converting the vector of an examinees scores at each occasion (X_{it}) into a final score (X_{ifinal}):

In North Carolina, a student's final score was the highest of the two retests. Therefore,

an examinee's final score after one retest is $max(X_{i1}, X_{i2})$.

I repeat this simulation 10,000 times, with each new simulation having a new conditional θ sample and a new error draw. I do not draw an initial error term, because this initial error is the residual between the sampled true score and the initial observed score. In each simulation, I estimate the final pass rate and the mean score increase by initial test score. The expected value for both these quantities is the mean across simulations. For the mean score increase, the final point estimate is a mean across simulations of means within each simulation.

The 95% prediction interval is the values between the 2.5% and 97.5% percentile across the simulation results. If the quantities of interest in the actual exam are outside the prediction interval, we can reject at the .05 level that the changes in the quantity of interest were only due to test measurement error. In this example, I do not need to account for sampling error because the initial observed score is fixed; drawing a new true score sample accounts for measurement error in the initial score. A researcher looking to also account for sampling error should redraw a new observed score distribution in each simulation.

This model makes a few core assumptions about measurement error for the sake of simplicity, but all of these assumptions are specific to this example and not endemic to the broader analytical method. First, I assume the internal consistency reliabilities in the exam technical documentation are an accurate proxy for test-retest reliability. In practice, test-retest reliabilities are lower than internal consistency reliabilities, and these estimates of the retesting effect should be taken as a conservative estimate of how big the retesting effect is. I also explicitly test the results' sensitivity to this assumption by running the simulations for range of reliability values. This also implicitly tests the models sensitivity to misestimations of σ_X^2 because σ_ϵ^2 is a multiplicative combination of σ_X^2 and $1 - \rho$. Therefore decreasing ρ has the same effect as increasing σ_X^2.

Second, I assume the error is homoskedastic across testing occasions and students. Both the initial test and the retest were exams drawn from the same item bank and were designed to support comparable scores, leaving no compelling reason to expect the error distribution

to be different across the different testing occasions. Similarly, the technical documentation provided by the North Carolina Department of Public Instruction shows little variance in the exam error either by grade level or score level, which supports assuming homoskedasticity across students.

In general heteroskedacity will have little impact on the magnitude of the retesting effect on averages. The retesting effect goes up with an increase in overall error but shifting the error between occasions or examinees does little to the net effect. This is because while lower reliability for specific examinees or occasions will increase the retesting effect it will be averaged out by an decrease in the effect for the higher reliability occasions or examinees. This does not apply to the pass rate, where heteroskedacity across students should change the retesting effect. The direction of this effect is a little complicated, however in general if the test error variance is larger farther away from the minimum passing score then the retesting effect is larger. This is because near the pass/fail point only a little error variance is required to change a student's pass/fail status and father away from the pass/fail point more error variance is required to flip a student to passing. On the other hand students too far from passing will very rarely have their status flipped and so there is an ideal distribution of error across students that will maximize the retesting effect on the pass rate.

Third, I assume that, conditional on a student's observed score, a student's decision to retest in independent of their true score. As previously noted, 98% of Level II students retook the exam, and this high compliance suggests little room for the non-response bias that would make the independence assumption problematic. If students with high true scores are more likely to retest then retesting effects will be larger than estimated in this model because of mean reversion. It is also important to note, this model is not able to disentangle true change in proficiency from selection into retesting, which can more significantly alter the interpretation of results in cases with lower retesting rates. Again, the retesting independence assumption means the estimates of the retesting effects in this example are conservative.

1.2.4 North Carolina Retesting Results

Figure 1.2 shows the effect of retesting on just the distribution of retested students. Initially, all the retested students have failing scores, but after the retest, some density has moved to the right of the pass/fail line. There is a discontinuity at the pass point, but when we limit the data to just the retested students, more students have scores just below passing than just above passing. This happens because the final score is the maximum of the two testing occasions. At one point below passing, the final score histogram contains all final observations where the initial score was one point below passing and the second score was more than one point below passing, and all final observations where the initial score was more than one point below passing and the second score increased to be only one point below passing. In contrast, the density at zero only contains observations where the second score lands at the minimum passing score.

The retesting effect alone causes 13% of the initially failing test-takers to pass with the final score and increases the mean score by .1 standard deviations (1 scaled score point) for the retested students. These results occur using only modeled data, and so both the change in the pass rate and the increase in the mean score are statistical artifacts of the retesting policy and are not caused by teacher or student behavior.

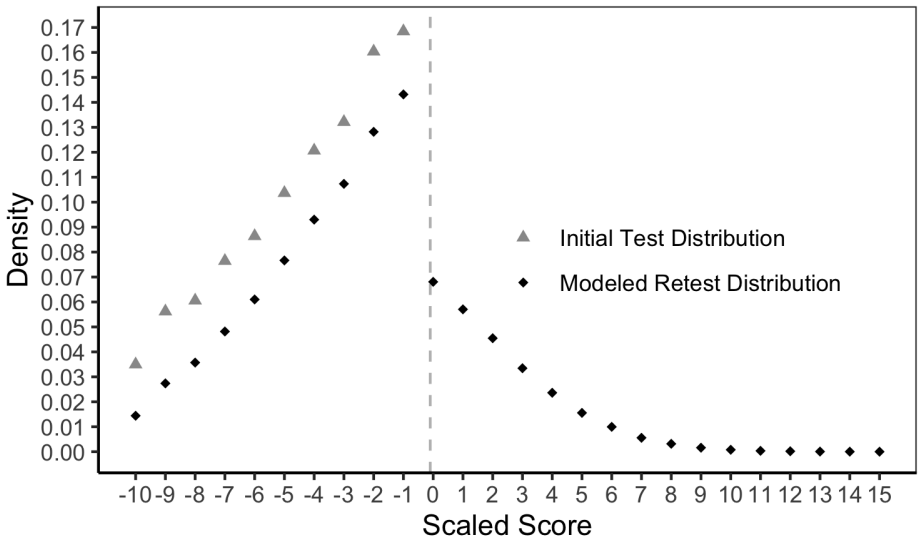


Figure 1.2: *Initial and retested score distributions for retested examinees for an exam with a reliability of .92. The dotted line represents the pass/fail line.*

Figure 1.3 shows the magnitude of the retesting effect on students who initially scored one point below passing. As expected, the retesting effect decreases as the exam reliability increases. The retesting effect operates through measurement error, and so exams with less measurement error will see less change between exam occasions. What is noticeable is that even in exams with very high reliabilities, a single retest significantly distorts the distribution. At an implausibly high reliability of .99, the retesting effect causes 38% of students who initially fail the exam by one point to pass on the retest and causes their scores to go up by an average of .06σ (i.e., .6 scaled score points). At a more typical reliability of .92, the retesting effect causes 51% of students initially one point from passing to now pass and creates an average score increase of .19σ (i.e., 1.9 scaled score points).

The solid vertical bars at each point on the graphs are 95% prediction intervals of the retesting effect. In this case, the sample size is large, and so the intervals are small across all reliability values. Even so, the prediction intervals are narrowest for lower reliability values because less measurement error leads to less uncertainty in the final distribution. Notably, the lengths of the prediction intervals are small compared to the range of estimates across assumed reliabilities. This implies that imprecision in the reliability estimate has a bigger

impact on the retesting effect estimate than classical uncertainty.

Accounting for random variation in measurement error and imprecision in the reliability estimate, this modeling still consistently shows growth in the underlying pass rate. However, imprecision in the reliability estimate makes it unclear whether the change in the underlying mean test score across testing occasions was above zero. The dotted horizontal lines on the graphs represent the retest pass rate and mean score increase for the actual North Carolina students who initially scored one point from the pass cut. The unadjusted pass rate for students initially one point from passing was 63.2%, which is above the prediction interval for all reliability values. The unadjusted mean test score growth was .17σ, which is just below the prediction interval at the assumed reliabilities of .9 and .92 where the prediction intervals are (.213,.217) and (0.191,.187) and just above the prediction interval at the assumed reliability of .94 where the predication interval is (0.162, 0.158). Regardless of the true test reliability value, it is clear adjusting for the retesting effect is important because the mean test score growth estimate drops by between 95% and 128% when adjusted for retesting.

ART calculate that, given what is known about test score fade out, their results imply a difference of .06σ at the point of retesting between students who initially narrowly failed and students who initially narrowly passed. A score difference of this magnitude is unlikely to have resulted only from test score growth in the retested group. The model only estimates that the retested group had an adjusted test score growth between testing occasions larger than .06σ for assumed reliabilities of .97 or above. It is unlikely that the test had a reliability as high as .97; less than 2% of U.S. state math accountability exams have reported reliabilities above .96 (Ho and Reardon, 2015). A gap of .06σ between the retested and non-retested students, therefore, implies a decline in the average math proficiency of the non-retested students in the time between the testing occasions.

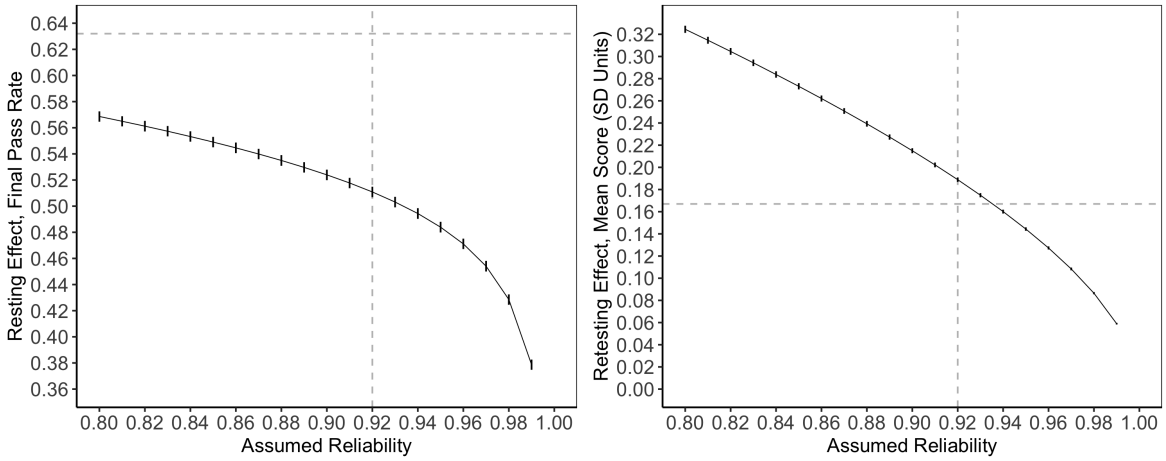


Figure 1.3: *The retesting effect on the pass rate (left) and mean score (right) for students who initially scored one point below passing calculated for a range of reliabilities. The vertical solid lines at each point represent 95% prediction intervals. The vertical dotted line is at the test reported reliability of .92 and the horizontal dotted line is at the empirical values observed in the data.*

One advantage of a descriptive analysis enabled by adjusting for the retesting effect over only performing a regression discontinuity analysis is the ability to examine students across the full range of failing scores and not just students at the pass/fail threshold. Each point in Figure 1.4 is the empirical final pass rate or average score increase minus the pass rate or average score increase from the retesting simulations. The adjusted gains in the pass rate were largest for students who initially scored between four and six points below the minimum pass score (i.e., about 15% for each group of students). However, the adjusted gains in the average test score were largest, $.06\sigma$ (.6 scaled score points), for students ten points below the pass point (i.e., the lowest scoring students in the sample) and mostly decreased as the initial student score increased. The only exception is mean test score growth was larger, or the test score losses were smaller, for students who initially were one or two points below passing than students who were initially three points below passing.

This is suggestive evidence that the retesting policy most advantaged the lowest performing students and not “bubble students" nearest the passing margin. This is in contrast to prior research on test based accountability policies in other contexts, which found that such systems push test score gains towards students nearest the passing margin (Burgess,

Propper, Slater, and Wilson, 2005; Neal and Schanzenbach, 2010; Reback, 2008; Springer, 2008). One reason the results here may be different is that the North Carolina accountability formula took into account both the percentage of passing students and growth in the average test score, not just the percentage of passing students.

Overall, none of the analysis in this example contradicts the research done by ART, but it does provide additional context for that research. When retesting effects are adjusted for, there was near zero test score growth across testing occasions for the initially failing students on the pass/fail margin. If the test score differences between the retested and non-retested were driven by score differences at the point of retesting, the near zero mean test score growth between testing occasions for the retested students suggests that the non-retested students had some proficiency loss between testing occasions. To the extent this generalizes to schools outside North Carolina, it indicates that there are educational costs to having state accountability exams too many weeks before the end of schools, and effort should be placed on engaging students between exams and the last day of school. This analysis also shows that while the North Carolina accountability policies arbitrarily benefitted retested students versus non-retested students on the margin, within the retesting window, lower performing students had more test score gains than higher performing students. The gains made by the lower performing students are large in the context of education effect sizes and part of the intentional goals of test based accountability systems.

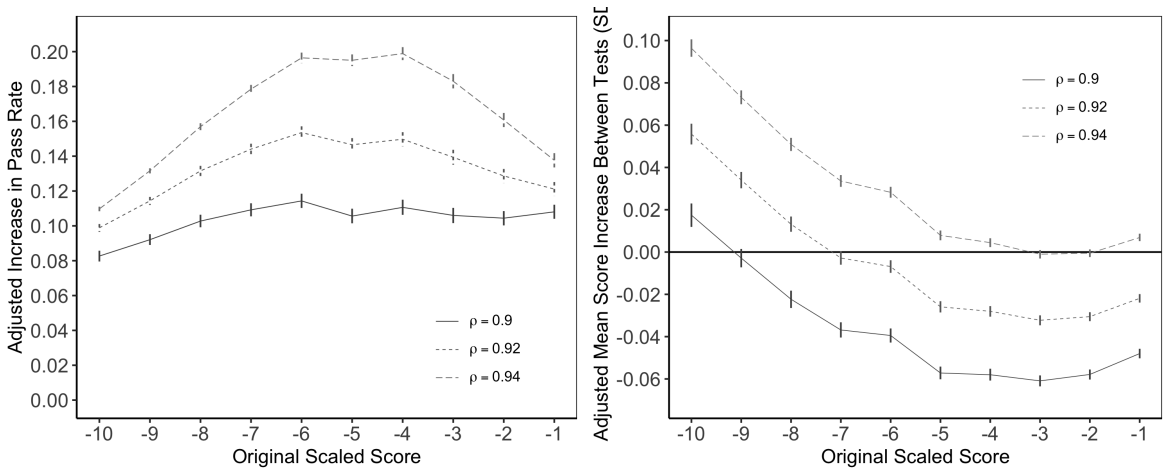


Figure 1.4: *The adjusted increase in the pass rate (left) and mean score (right) between retests for North Carolina students assuming a reliability of .90,.92, and .94.*

1.3 Rescoring on the New York Regent Exam

1.3.1 Analytical Rescoring Model

As with the retesting example, I use a CTT based model for rescoring. Here, the observed score is modeled as an additive combination of a student's true proficiency, the measurement error from the testing occasion, and the measurement error from the exam grading. Thus the observed score for examinee i at grading occasion g is:

$$X_{ig} = \theta_i + \epsilon_{test,i} + \epsilon_{grade,ig} \tag{1.5}$$

$$\epsilon_{test,i} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_{test,i}}),\ \epsilon_{grade,ig} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_{grade,ig}})$$

$$Cov(\theta_i,\ \epsilon_{test,i}) = 0, Cov(\theta_i,\ \epsilon_{grade,ig}) = 0,\ Cov(\epsilon_{testi}, \epsilon_{grade,ig}) = 0, \quad \text{for all } i \text{ and } g$$

where θ_i is the true proficiency of examinee i, $\epsilon_{test,i}$ is the testing occasion measurement error for examinee i, and $\epsilon_{grade,ig}$ is the grading error for examinee i at grading occasion g.

Unlike the retesting example, the test occasion measurement error and the grading error are broken out into two separate model terms. The test occasion error is all the error from the actual taking of the exam (e.g., how well the examinee slept the night before,

the presence of distractions in the test-taking room, etc.) and is fixed once the test taker completes their test. Grading error is all error introduced in the grading process (e.g., an unusually stringent or lax grader, the presence of distractions in the grading environment, etc.). Rescoring only involves one testing occasion, so the testing occasion error is fixed for each test taker. As in the retesting example, I assume a fixed θ for each examinee, but unlike with retesting this assumption is true by construction because test takers cannot improve their true proficiency between grading occasions.

Analogous to selective retesting, selective rescoring distorts the distribution because rescoring redraws $\epsilon_{grade,ig}$ for only some test takers. Modeling rescoring with simulation requires one more input than the retesting model, the share of the error due to grading, not occasions. Therefore the five inputs to the rescoring model are: the testing error variance ($\sigma^2_{\epsilon_{testi}}$), the grading error variance ($\sigma^2_{\epsilon_{grade,ig}}$), a θ distribution, a rule for determining when an examinee is rescored, and a rule for converting the vector of an examinees scores at each grading occasion (X_{ig}) into a final score (X_{ifinal}). As with retesting, I define the rescoring effect as the difference between the final modeled distribution (X_{final}) and the initial distribution (X_1).Given that θ can't change between scoring occasions, all observed changes in the distribution between scoring occasions outside the predication interval indicate a problem in the scoring process.

1.3.2 New York Regent Exam Background

The New York Regent exam is the longest running and one of the largest high school exit exam programs in the United States. High school students in New York State must score at least a scaled score of 65 on exams in each of five core subjects (English, Geography, Mathematics, Science, and US History) to receive a "Regents Diploma". Students who entered high school before 2008 and who scored at least a 55 in each of the core subjects were eligible for a lesser credential, known as a "Local Diploma".

All Regent exams contain both multiple choice and constructed response sections. The multiple choice questions are graded by machine with only one possible correct answer.

However, the constructed response portion is graded by human graders and requires the grader to exercise judgment in their grading. Therefore there is an opportunity for error in the grading of the constructed response portion of the Regent Exam.

In this example, I re-examine the controversial pre-2011 rescoring policy, where exams initially between 60 and 64 were rescored. Before 2011 schools could also choose to rescore tests between 50 and 54. I do not focus on the 50 to 54 rescoring window because the majority of schools were not rescoring at the lower stakes 55 threshold and there is no clear record of which schools were rescoring between 50 and 54 (Dee et al., 2019; New York State Education Department, 2009a). I do, however, account for the 50 to 54 rescoring window in my analysis.

In this study, I re-analyze the rescoring policy using data from the June 2009 sitting of the Integrated Algebra Regent Exam. The data is fully de-identified and only contains the scaled score for the 79,166 students who took the June 2009 Integrated Algebra Exam in New York City. The scores in the data set are post any rescoring that was done and combines each student's multiple choice and constructed response scores. One reason I use the June 2009 Integrated Algebra Regent Exam is the New York Regents did a technical audit of this exam which provides disaggregated standard deviations and the reliabilities of the constructed response and multiple choice questions (New York State Education Department, 2009b), which I use in my analysis.

All analysis for this example is performed on the raw scores and not the scale scores. However, the data is reported in scaled scores and not raw scores. Raw scores are obtained using the Regent raw score to scaled score conversion chart which is based on a one-parameter IRT model, so each raw score only maps to a single possible scaled score. In some cases, multiple raw scores map to a single scaled score, or a given scaled score may have no corresponding raw score. I assume uniform density across the possible raw scores in cases where a scaled score maps to more than one raw score. There are no cases of this near the 55 or 65 scaled score cut points, although a scaled score of 63 does not map to any raw score and does not appear in the data.

1.3.3 New York Rescoring Model

In the preceding retesting example, I had the test score distribution from both the original test and the retest, but only for students in the retesting window. Here I have the reverse, I only have data from the final rescored distribution, but I have the distribution for all students and not just those in the rescoring window. I demonstrate how either case provides enough data to implement this method. I use the final rescored empirical data from the Integrated Algebra exam to model an unrescored smooth distribution similar to the New York Regent Exam's empirical distribution. Based on the shape of the data, I treat the data as a log normal distribution. I estimate the unrescored distribution by dropping the data from the two rescoring regions (50-54 and 60-64) and at scores within two points of the rescoring regions (55,56 and 65,66). I also drop all data within ten raw score points of the minimum and maximum scores to prevent the tails from having undue influence on the estimation. I use this data to estimate the uncensored unrescored observed score distribution using the fitdistrplus package in R, which fits censored distributions using non-parametric maximum likelihood estimation Delignette-Muller and Dutang (2015). Using this process, the estimated smooth distribution on the log scale has a mean of 3.44 and a standard deviation of .51 in the raw score.

As with the retesting example, I impose a homoskedasticity assumption for simplification. Therefore the observed score for each test taker i at grading occasion g is modeled as:

$$log(X_{ig}) = \theta_i + \epsilon_{test,i} + \epsilon_{grade,ig} \tag{1.6}$$

$$\epsilon_{test,i} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_{test}})$$

$$\epsilon_{grade,ig} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_{grade}})$$

$$Cov(\theta_i, \epsilon_{test,i}) = 0, Cov(\theta_i, \epsilon_{grade,ig}) = 0, Cov(\epsilon_{test,i}, \epsilon_{grade,ig}) = 0, \text{for all } i \text{ and } g$$

As explained above, this rescoring model takes five inputs: the testing error variance ($\sigma^2_{\epsilon_{test}}$), the grading error variance ($\sigma^2_{\epsilon_{grade}}$), a θ distribution, a rule for determining when an examinee is rescored, and a rule for converting the vector of an examinees scores at each grading

occasion (X_{ig}) into a final score (X_{ifinal}).

1. Estimate of the testing error variance ($\sigma^2_{\epsilon_{test}}$) and 2. the grading error variance ($\sigma^2_{\epsilon_{grade}}$):

The overall error variance is modeled by reliability adjusting the variance estimate for the initial unrescored observed score distribution where $\sigma^2_{log(X)}$ is the variance of log(X) and ρ is the test reliability.

$$\sigma^2_{\epsilon} = (1-\rho)\sigma^2_{log(X)} \tag{1.7}$$

Using the empirical data $\sigma_{log(X)}$ is estimated as .51 raw score points. The technical documentation reports the test reliability as .93 New York State Education Department (2009a). This reliability is from the non-logged score, but the correlations between parallel scores in the logged scale and the non-logged scale are close. Therefore the reliability in the non-logged scale provides a close approximation of the reliability in the log scale. As in the retesting case, I also run the model for all other reliabilities between .75 and .99 in .01 increments as a sensitivity check.

I break this error variance estimate into its testing and grading components by assuming that the proportion of the error attributable to grading is the same as the proportion of overall test score variance attributable to the constructed response questions. This assumption is non-trivial, however in practice, I run the model over a range of reliabilities, which implicitly tests a range of values for $\sigma^2_{\epsilon_{grade}}$. According to the technical audit of the Integrated Algebra exam, the constructed response questions account for 23% of the test score variance New York State Education Department (2009b). This means the error variances are modeled as:

$$\sigma^2_{\epsilon_{test}} = .77(1-\rho)\sigma^2_{log(X)} \tag{1.8}$$

$$\sigma^2_{\epsilon_{grade}} = .23(1-\rho)\sigma^2_{log(X)} \tag{1.9}$$

3. Estimate of the θ distribution:

The true score distribution is modeled as:

$$\theta_i \overset{iid}{\sim} N(\mu_{log(X_{i1})}, \sigma^2_{\theta})$$

where $\mu_{log(X_{i1})} = 3.44$.

The true score variance is modeled by reliability adjusting the variance estimate for the initial unrescored observed score distribution where $\sigma^2_{log(X)}$ is the variance of log(X) and ρ is the test reliability Lord and Novick (1968).

$$\sigma^2_\theta = \rho\sigma^2_{log(X)} \tag{1.10}$$

4. A rule for determining which examinees are rescored:

The initial observed score is estimated by drawing 79,166 observations from each of the error distributions and the true score distribution. The log of the initial observed score is the sum of these terms, such that:

$$log(X_{i1}) = \theta_i + \epsilon_{test,i} + \epsilon_{grade,i1} \tag{1.11}$$

Under the New York Regent rescoring policy, a student was rescored if $26 \leq X_{i1} \leq 29$, where 26 is the raw score that corresponds to a scaled score of 60 and 29 is the raw score that corresponds to a scaled score of 64. When the initial score is in this rescoring region, the grading error is resampled and the regraded score is estimated as:

$$log(X_{i2}) = \theta_i + \epsilon_{testi} + \epsilon_{gradei2} \tag{1.12}$$

5. A rule for converting the vector of an examinees scores at each grading occasion (X_{ig}) into a final score (X_{ifinal}):

The New York Regent exam provided no clear guidance to teachers on how to do the rescoring (New York State Education Department, 2009b), so I model four potential scoring rules to see which one most matches the empirical distribution. 1) a policy where a student is assigned their new score if it is higher than their original score (Highest Score); 2) a policy where rescored students always receive their second score (2nd Score); 3) a policy where a student is only assigned their new score if it moves the student from a non-passing to a passing score (Passing Score); 4) a policy where if the regrade moves the student from a non-passing to a passing score they are assigned the minimum passing score (Minimum Passing Score).

Each simulation is repeated 10,000 times and has a true score draw, a new testing error draw, and two new grading error draws. In each simulation, I estimate four quantities of interest: 1) The pass rate conditional on initially being in the rescoring region (i.e., initially scoring between 60 and 64); 2) the overall pass rate; 3) the pass rate for students with true scores below the cut point (i.e., the false pass rate); and 4) the failure rate for students with true scores above the cut point (i.e., the false fail rate). This first quantity will provide a basis for our score scrubbing test. The pass rate conditional on initially being in the rescoring region is the amount of mass we see in the rescoring region in the unrescored distribution minus the mass we see in the rescoring region in the rescored distribution. If the amount of mass missing from the rescoring region in our empirical distribution is outside the 95% prediction interval for the amount of mass missing in our simulated rescoring distribution, then we can reject the hypothesis at the .05 level that empirical distribution came from an unbiased rescoring process.

One thing missing from the original conversation about the Regent rescoring policy was an evaluation of the merits of the policy as written. In my analysis, I compare the Regent rescoring policy to two other potential scoring policies: 1) not rescoring and leaving the cut point at 65 and 2) not rescoring and moving the cut point to 64. I compare the rescoring policy to these policies by estimating the pass rate, false pass rate, and false fail rate for each of these policies in each of the simulations.

The assumptions in this example are similar to those in the retesting example. First, I assume homoskedasticity across grading occasions and test takers. In this example, we might expect some heteroskedasticity across grading occasions because teachers were aware they were regrading exams, and so it's possible they were more careful at the second grading occasion than they were on the first grading occasion. However, like in the retesting example, the results should be primarily dependent on the total error and not how error is distributed across occasions. Heteroskedasticity across test takers has a greater ability to change the results, but the technical documentation shows similar conditional standard error of measurement values for all test takers in the rescoring region (New York State

Education Department, 2009b). To the extent there is heteroskedasticity across test takers, the errors are a little larger farther away from the cutpoint, and so by the same reasoning used in the retesting example, assuming homoskedasticity produces a conservative estimate of the rescoring effect.

Second, I assume the exam's reported reliability is a reasonable proxy for the reliability across testing and grading occasions. This assumption about reliability is closely related to the third big assumption in the model; that the constructed response variance is a good proxy for the proportion of error variance attributable to grading. The model's sensitivity to both these assumptions is tested by running the model for a range of reliabilities. Increasing the test reliability reduces the grading error variance and, therefore, shows the model sensitivity to assumptions about the grading error's magnitude.

1.3.4 Rescoring Results

A graph of the empirical test score distribution shows there is indeed a large discontinuity in the empirical distribution at the pass point; there is more than six times the number of students who received the minimum passing score on the exam than students who received the maximum failing score (Figure 1.5). As expected, in the rescoring region, the empirical distribution and the estimated smooth distribution do not closely match. However, outside the rescoring region, the estimated smooth distribution provides a reasonable estimate of the empirical distribution.

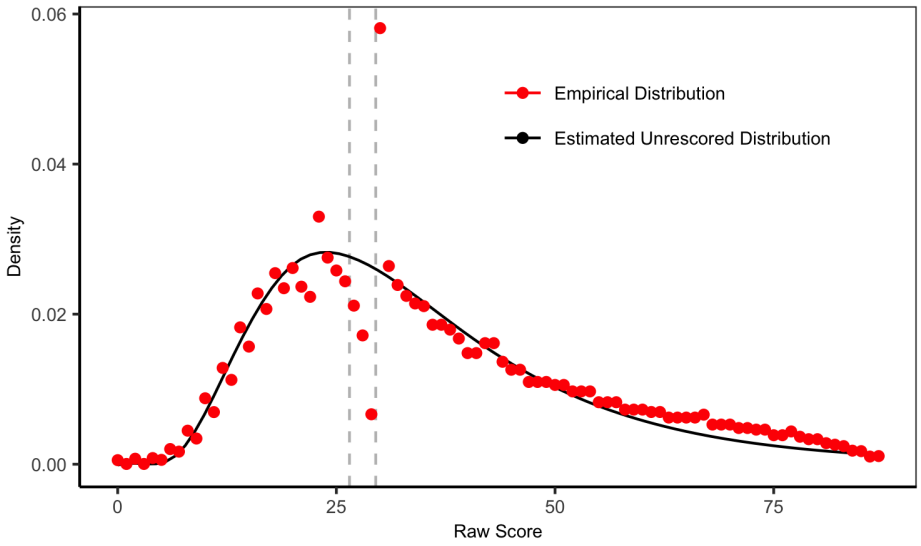


Figure 1.5: *Estimated unrescored distribution compared to the empirical NY Regent Integrated Algebra distribution. The left dotted line is the lowest score a student needs to be rescored, and the right dotted line is the lowest passing score.*

Graphs of the distribution under each of the proposed scoring rules show that the Minimum Passing Score scoring rule (i.e., where students are rescored and then given a 65 if their new score puts them over the passing threshold) best explains the empirical distribution (Figure 1.6). The distribution pattern for each of the other scoring rules matches the empirical data less well than the Minimum Passing Score scoring rule. In the Second Score scoring rule model, density in the rescoring region is dispersed out of the rescoring region to both the left and right of the rescoring window. This leaves little density left in the rescoring region and creates discontinuities to both the left and right of the rescoring region, where the empirical distribution only has a discontinuity to the rescoring region's right. In the Highest Score model, density is only shifted to the right, as students are moved to a higher score through the rescoring process, and so there is only a discontinuity at the pass threshold, as in the empirical data. However, the Highest Score rule model has the fewest students at the minimum score needed to qualify for rescoring, as the policy shifts all rescoring students to the right, even within the scoring window. This leads to an upward sloping pattern even within the rescoring region, which again is not what we see in the

empirical distribution. Under the Passing Score policy, students are again only shifted to the right, but only if the rescoring changes their pass/fail status. This means that students are not shifted within the rescoring window. Students with scores farther away from passing are less likely to benefit from the policy, and so there is less change to the proportion of students at points farther away from the pass point. This leads to a downward sloping pattern in the rescoring range and the least students at the maximum failing score, both of which are consistent with the empirical data. However, the Passing Score policy has less density at the minimum passing score than the empirical distribution but more density at passing scores just above the minimum than the empirical distribution. The Minimum Passing Score scoring rule has the same data pattern in the rescoring window as the Passing Score scoring rule but concentrates all the excess density over the pass point on the minimum passing score, just like the empirical data.

A rescoring rule that only changes students' scores if they go from failing to passing and then only gives them the minimum passing score is not "scrubbing" in the traditional sense, but it does not represent psychometric best practices either. To the extent to which the test was just focused on providing a pass/fail determination, this scoring rule provides the same decision as a more traditional rescoring rule, like the Highest Score procedure. However, to the extent the exam was serving other purposes (e.g., providing information to students and schools about student academic proficiency, allowing students to place out of remediation in the CUNY system), the Minimum Pass Score process does not provide a good estimate of student scores.

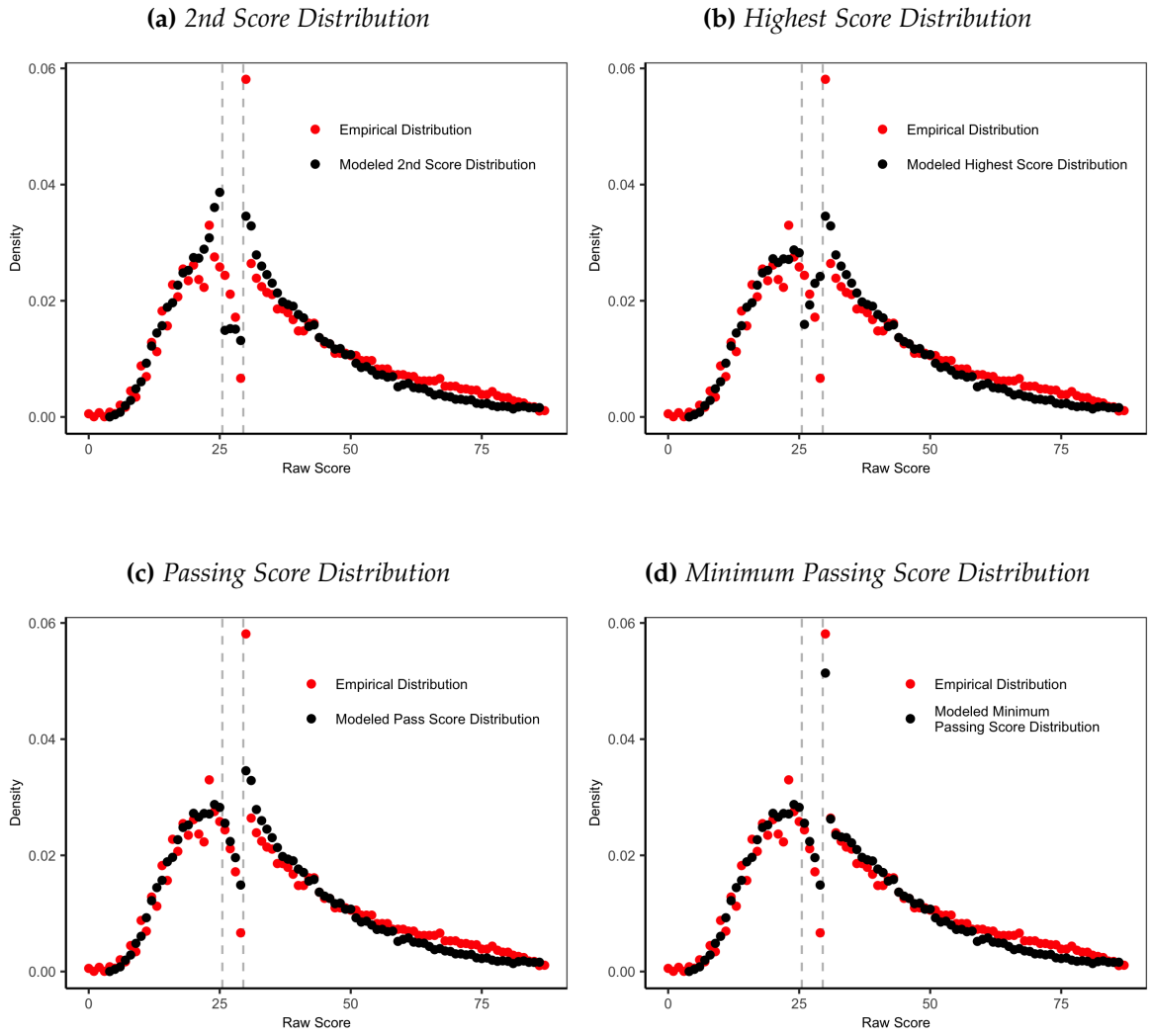


Figure 1.6: *Comparison of the empirical distribution to the modeled re-scoring policies using an assumed reliability of .93. The left dotted line is the lowest score a student needs to be re-scored and the right dotted line is the lowest passing score. The modeled scoring policies moving left to right are as follows; students with a scaled score between 60-64 are re-scored and: 1) given the second of the two scores, 2) given the highest of the two scores, 3) given the second score if the second score changes the student from failing to passing, 4) given the minimum passing score if the second score changes the student from failing to passing*

I also find that rescoring accounts for 67% of the missing density in the rescoring region (Figure 1.7) at the exam reliability of .93 reported by the technical documentation. That is, conditional on initially being in the rescoring region, the rescoring policy alone will cause 24% of students to end up with a passing score. This is consistent across all rescoring

policies because all policies have the same effect on the number of rescored students who end up passing; the different policies only have different implications for the distribution's shape. The 95% prediction interval for the pass rate conditional on being in the rescoring region is (22%, 26%). In the empirical data, we see a conditional pass rate of 36%, which is outside the prediction interval, and we can reject a hypothesis of no scrubbing at the .05 level (Figure 1.7). However, I still find that not accounting for rescoring significantly overestimates the amount of scrubbing. On the 2009 Integrated Algebra Exam, not accounting for the bias caused by the rescoring policy produces scrubbing estimates three times too large.

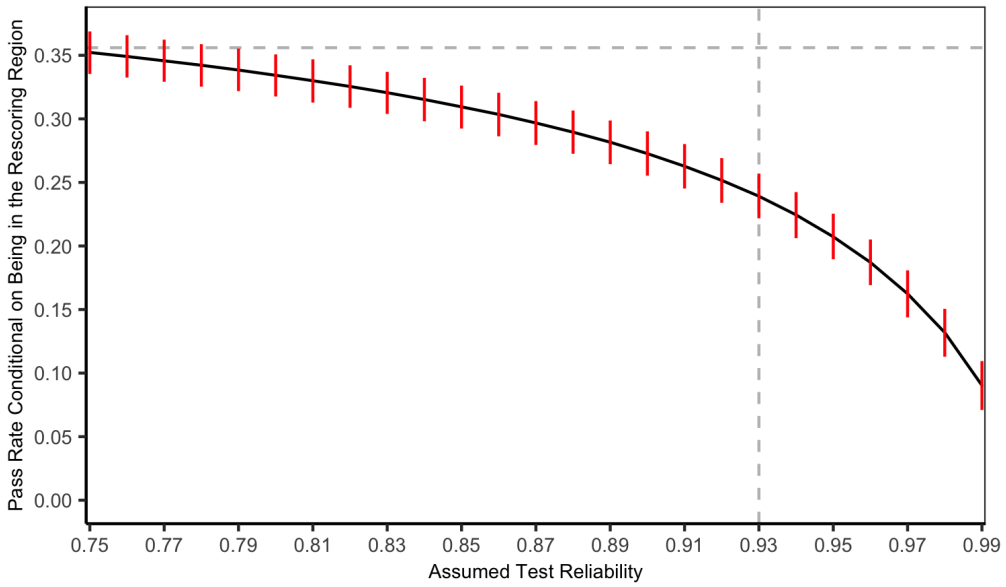


Figure 1.7: *The final pass rate conditional on initially being in the rescoring region. The dotted horizontal line is the conditional pass rate for the empirical distribution. The dotted vertical line is at a reliability of .93, which is the exam reported reliability.*

Comparing rescoring to lowering the pass cut by one point shows the two policies to be similar (Figures 1.8 and 1.9). All else being equal, when I model the Regent Exam with no rescoring and pass cut at 65, the pass rate is lower, the false pass rate is lower, and the false fail rate is higher than for the modeled distribution with rescoring. However, at conventional test reliabilities, the pass rate, false pass rate, and the false fail rate are almost the same for the rescoring model and the model of lowering the pass cut by one point; where the false

pass rate is defined as passing when the test taker has a true score below 65 and where the false fail rate is defined as failing when the test taker has a true score above 65.

One potential implication is that the Regent Exam possibly could have lowered their pass cut by a point, instead of rescoring, since it has the same practical impact as rescoring but is easier and cheaper to implement. On the other hand, a rescoring policy sends a different message to students then lowering the pass cut. Rescoring may better communicate that the proficiency to aim for is 65 while also putting extra weight on reducing false failures. Whether rescoring or lowering the pass threshold is the better policy, it is unclear that the Regent exam administrators understood they were practically equivalent, and the New York Regent Board did not communicate that to the public in the Wall Street Journal article's aftermath. Other exams with rescoring policies should compare the effect of rescoring to the effect of lowering the minimum passing score and evaluate whether the exam goals are better served by lowering the cut score, which is easier and cheaper to administer.

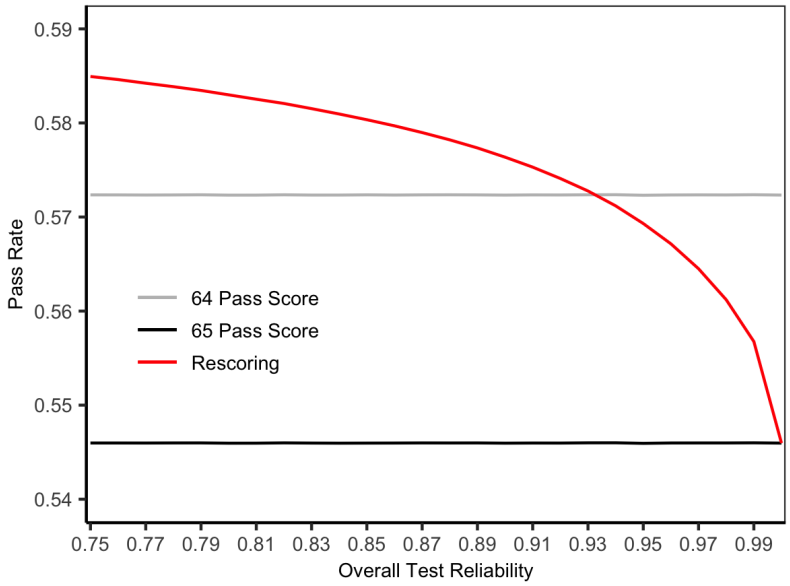


Figure 1.8: *The modeled overall pass rate under different scoring policies for a range of reliabilities.*

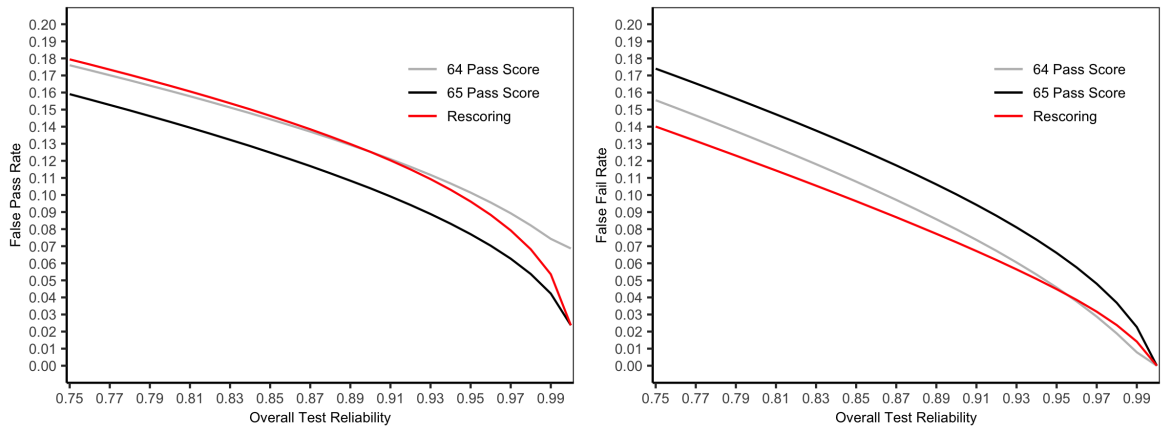


Figure 1.9: *Estimates of the false pass rate and false fail rate under each policy for a range of reliabilities. The left panel shows the percentage of students with true scores above 65 that fail under each policy. The right panel shows the percentage of students with true scores below 65 that pass under each policy.*

1.4 Conclusion

Testing policy is inextricably linked to how test scores must be analyzed. In this paper, I take a common testing policy, selectively rescoring and retesting students, and show how this policy creates statistical mirages that can cause researchers and policymakers to see human behavior that isn't there. More importantly, in this paper, I develop a method for dispelling these mirages. I demonstrate how a combination of Classical Test Theory and simulation can be used to estimate the direct statistical effect of selective rescoring and retesting policies on the whole test score distribution and therefore adjust any distribution level quantity to account for rescoring and retesting.

I give two examples where adjusting for rescoring and retesting changes the data interpretation. In North Carolina, students who initially failed the standardized exam by one point had an unadjusted mean test score growth of $.17\sigma$ between testing occasions. However, adjusting for the statistical effect of retesting, this growth shrunk to near zero. On the New York Regent, the unadjusted score scrubbing estimate showed three times as much scrubbing as the estimate that adjusted for rescoring.

The rescoring and retesting adjustment method in this paper is flexible. This method can be used in test development, test policy decisions, and secondary data analysis. Unlike prior research, this method allows policymakers and researchers to estimate any distribution level quantity of interest. Additionally, this adjustment method only requires a limited amount of data. I show how to do calculations when data is either missing outside the retesting interval or when the initial score data is missing.

The method in this paper is also easily adaptable to contexts outside this paper. Here, I make assumptions about the amount of grading and testing error variance, the consistency of the error variance across testing and grading occasions, and which students are rescored or retested. All of these assumptions are easily changed to match the context of the analysis. Because rescoring and retesting policies are common, there are many other settings where these methods could be used. To give just one example, measurement experts overseeing the administration of legal bar admittance exams have expressed concern that selective rescoring exams must inherently be biased because the graders know the examinees initially failed (Albanese, 2016). However, the bar scoring procedure's robustness is an empirical question that this study provides the tools to answer.

This research attempts to bridge two literatures: psychometric literature focused on developing accurate test scores and economic literature interested in interpreting data to understand human behavior. In this study, I illustrate one method for using the tools psychometrics for doing economics style analysis. Future work leaves open the possibility of developing many more such methods.

Chapter 2

Characterizing Cross-Site Variation in Local Average Treatment Effects in Multisite RDD contexts with an Application to Massachusetts High School Exit Exam[1]

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 evaluating two methods for estimate cross-site treatment effect variance in multi-site RDDs. The first method combines a fixed intercepts/random coefficients (FIRC) model with a local linear RDD analysis. The second

[1]Co-authored with Luke Miratrix

method borrows techniques from random effects meta-analysis and employs them with the RDD model. We find that a restricted FIRC model works best if there is no variance in the running variable coefficients but is biased when there is variance in the running variable coefficients. When there is variance in the running variable coefficients, we find an unrestricted FIRC model works best if the average number of in bandwidth observations per site is less than 100, and the random effects meta-analysis model works best if there is more than 100 in bandwidth observations per site. We then 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, but that the treatment effect variance was consistently large enough for the treatment effect to have been negative in more than a third of high schools.

2.1 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 and Lemieux, 2008; Lee and 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, and Walters, 2013; Kling, Liebman, and Katz, 2007; Raudenbush and Bloom, 2015; Raudenbush, Reardon, and Nomi, 2012; Tipton, 2014; Weiss, Bloom, and Brock, 2014).

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 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, Bloom, Verbitsky-Savitz, Gupta, Vigil, and Cullinan, 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 multi-site 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 multi-site RDD. Raudenbush, Reardon, and Nomi (2012) in a study on statistical methods for multi-site instrumental variable analysis, estimate the cross-site variance of the first stage of a multi-site 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 multi-site study of early algebra in California middle schools. McEachin, Domina, and Penner (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 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 in an RDD FIRC model: 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, and 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, and Spiegelhalter, 2009).

We find that the restricted FIRC model works best as long as it isn't misspecified and there is no variance in the running variable coefficients. 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. We, therefore, 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 unrestricted FIRC model performs better than the random effects meta-analysis model when the average number of in bandwidth observations per site is not large. However, when the average number of in bandwidth observations per site is great than 100, the meta-analysis model has less bias and less error than the unrestricted FIRC 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.

2.2 Analytical Models

A linear RDD model frequently takes the following form:

$$\begin{gathered} Y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_2(Score_{ij} - Score_c) + \beta_3 T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\sim} N[0, \sigma_y^2] \end{gathered} \tag{2.1}$$

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 (Gelman and Imbens, 2019; Hahn, Todd, and Van der Klaauw, 2001; Imbens and Lemieux, 2008). 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 2.1). The first 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.

The other 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.

Model	Unpool	Partially Pool	Pool
Local Linear Regression (LLR)	β_0		β_1, β_2, β_3
Random Effects Meta-Analysis (Meta)	β_0, β_1, β_2, β_3		
Fixed Intercepts Random Coefficient One (FIRC One)	β_0	β_1	β_2, β_3
Fixed Intercepts Random Coefficient Two (FIRC Two)	β_0	β_1,β_2, β_3	

Table 2.1: *Coefficient Estimation in the Multi-Site RDD Models*

2.2.1 The Random-Effects Meta Analysis Model:

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

$$\begin{gathered} Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}(Score_{ij} - Score_c) + \beta_{3j}T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\sim} N[0, \sigma^2_{yj}] \end{gathered} \tag{2.2}$$

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

$$\begin{gathered} VCov(\widehat{\beta_j}) = (X_j'X_j)^{-1}\widehat{\sigma_j}^2 \\ \widehat{\sigma_j^2} = \frac{1}{n_j - 3 - 1}\sum(Y_{ij} - \widehat{Y_{ij}})^2 \end{gathered} \tag{2.3}$$

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:

$$\begin{aligned} VCov(\widehat{\beta_j}) &= (X_j' X_j)^{-1} \overline{\widehat{\sigma_j}^2} \\ \overline{\widehat{\sigma_j^2}} &= \frac{1}{\sum n_j} \sum (n_j \hat{\sigma}_j^2) \end{aligned} \tag{2.4}$$

Consistent with the meta-analysis literature (DerSimonian and Laird, 1986; Higgins et al., 2009; Whitehead and Whitehead, 1991), 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:

$$\widehat{\beta_1} = \frac{\sum \widehat{\beta_{1j}} w_j}{\sum w_j},\ \widehat{SE_{\beta_1}} = \sqrt{\frac{1}{(\sum w_j)}} \tag{2.5}$$

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

$$\begin{aligned} \widehat{\sigma^2_{\beta 1}} &= max(0, \frac{Q - J - 1}{\sum \widehat{SE^{-2}_{\beta_{1j}}} - \frac{\sum \widehat{SE^{-4}_{\beta_{1j}}}}{\sum \widehat{SE^{-2}_{\beta 1j}}}}) \\ Q &= \sum W_j (\widehat{\beta_{1j}} - \frac{\sum W_j \widehat{\beta_{1j}}}{\sum W_j})^2, where\ W_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}}} \end{aligned} \tag{2.6}$$

where $\widehat{SE_{\beta_{1j}}}$ is the estimated site-level standard error for β_{1j} from Equation 2.4 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:

$$Q(\tau^2) = \sum_{j=1}^{J} \frac{(\widehat{\beta_{1j}} - \frac{\sum W_j \widehat{\beta_{1j}}}{\sum W_j})^2}{\widehat{SE^2_{\beta_{1j}}} + \tau^2}, where\ W_j = \frac{1}{\widehat{SE^2_{\beta_{1j}}}} \tag{2.7}$$

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 $Q(\tau^2) \leq \chi^2_{J-1}(1-\frac{\alpha}{2})$.

2.2.2 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.1, we also evaluate a partially restricted FIRC model with a partially pooled β_2 and a pooled β_3.

These two FIRC models are as follows:

FIRC One (restricted)

Level One - Observation:

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

Level Two - Site:

$$\beta_{1j} = \delta + e_{1j} \tag{2.8}$$
$$e_{1j} \overset{iid}{\sim} N(0, \sigma^2_{\beta_1})$$

FIRC Two (unrestricted)

Level One - Observation:

$$Y_{ij} = \alpha_j + \beta_{1j}T_{ij} + \beta_{2j}(Score_{ij} - Score_c) + \beta_{3j}T_{ij} * (Score_{ij} - Score_c) + \epsilon_{ij}$$
$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_y^2) \tag{2.9}$$

Level Two - Site:

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

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. We estimate both the models using restricted maximum likelihood (REML), however, we also present results using maximum likelihood in Appendix A.2.

We also formalize a method for estimating a confidence interval for the cross-site treatment effect variance estimate. In Appendix A.3, we test three possible methods for obtaining confidence intervals: 1) Wald standard errors, 2) Q-statistic inversion, and 3) profiled confidence intervals. Q-statistic inversion intervals worked best of these three methods, and therefore we use that method in our analysis.

The quantities in the Q-statistic are not estimated from the original multi-level model but from 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)

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_2 Score_{ij} + \beta_3 T_{ij} * Score_{ij} + \epsilon_{ij}$$

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

FIRC Two (unrestricted)

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j} Score_{ij} + \beta_{3j} T_{ij} * Score_{ij} + \epsilon_{ij}$$

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

The estimates from these OLS models are then plugged into equation 2.7, and the confidence interval is estimated the same way as in the Meta model.

2.3 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:

$$\begin{aligned} Y_{0ij} &= a_{0j} + b_{0j} Score_{ij} + \epsilon_{ij} \\ Y_{1ij} &= Y_{0j} + a_{1j} + b_{1j} Score_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_\epsilon^2) \\ Score_{ij} &= \mu_{Score} + r_j + \pi_{ij} \\ \pi_{ij} &\sim N(0, 1 - ICC_{Score}) \end{aligned} \tag{2.10}$$

where Y_{0ij} is the outcome absent treatment, Y_{1ij} is the outcome with treatment, σ_ϵ^2 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:

$$\begin{gathered} a_{0j} \sim N(\mu_{a0}, \sigma^2_{a0}), b_{0j} \sim N(\mu_{b0}, \sigma^2_{b0}), a_{1j} \sim N(\mu_{a1}, \sigma^2_{a1}), b_{1j} \sim N(\mu_{b1}, \sigma^2_{b1}) \\ r_j \sim N(0, ICC_{Score}) \\ n_j \sim Pois(\mu_n) \end{gathered} \tag{2.11}$$

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} 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 baseline value of .2. Across simulations, we vary the average observations per school (μ_n) from 10 to 350 with a baseline value of 130, the total schools (J) from 10 to 300 with a baseline value of 150.

We run all our simulations under conditions where the FIRC One model is correctly specified and σ_{b0} and σ_{b1} equal zero. We also run all our simulations under conditions when the FIRC One model is misspecified, and there is cross-site variance in the running variable

coefficients, where we set σ_{b0} and σ_{b1} equal to .05. In our evaluation of the LATE estimates we use σ_{b0} and σ_{b1} equal to .05 as the benchmark values. We also run simulations where all parameters are fixed at their benchmark values, σ_{b1}=.05, and the standard deviation of the control running variable coefficient (σ_{b0}) is varied from 0 to .3, and simulations where all parameters are fixed at their benchmark values, σ_{b0}=.05, and the standard deviation of the treatment running variable coefficient (σ_{b1}) is varied from 0 to .3.

Altogether, we run 52 simulations with each simulation parameter besides the parameter being varied fixed at its baseline value. Finally, while the overall average number of observations per site is directly manipulated across 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.

2.4 Simulation Results

2.4.1 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, 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 2.1). 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.

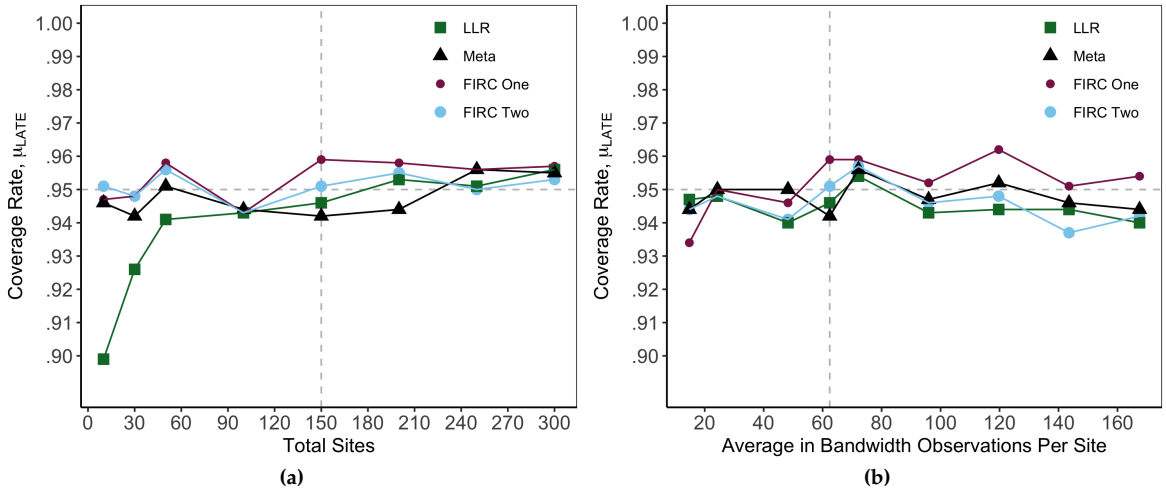


Figure 2.1: *The coverage rate of the local average treatment effect estimates across the local linear regression (LLR), the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random 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).*

2.4.2 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 2.2 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 2.3). 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 2.4). In Figure 2.4 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.

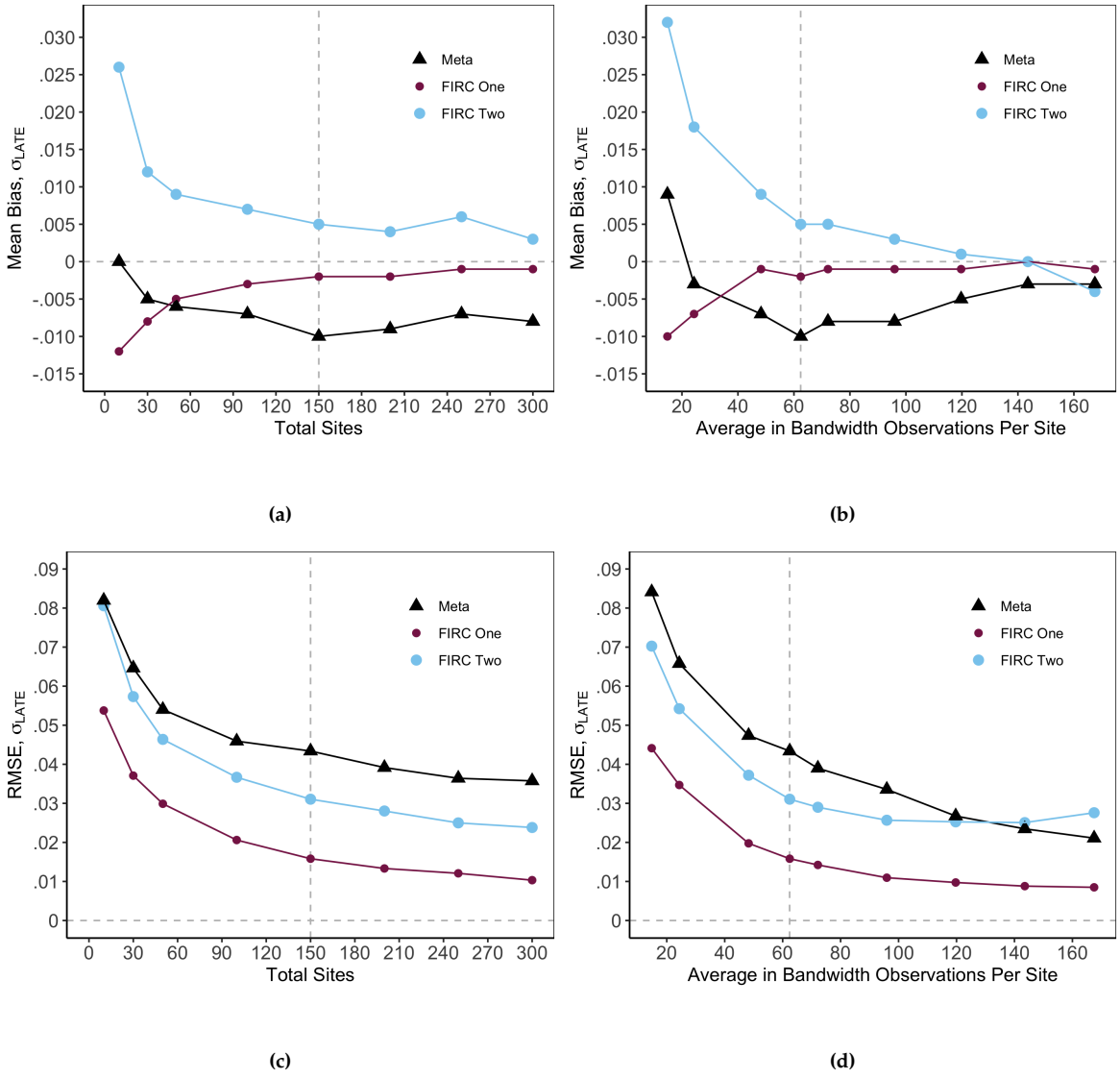


Figure 2.2: *The mean bias (top) and root mean squared error (bottom) in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients 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).*

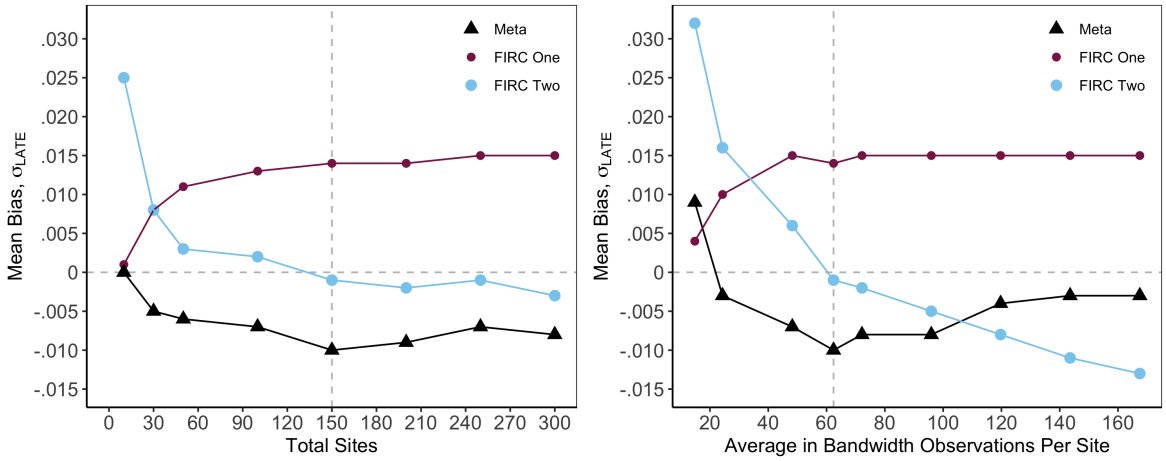


Figure 2.3: *The mean bias in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable coefficients (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).*

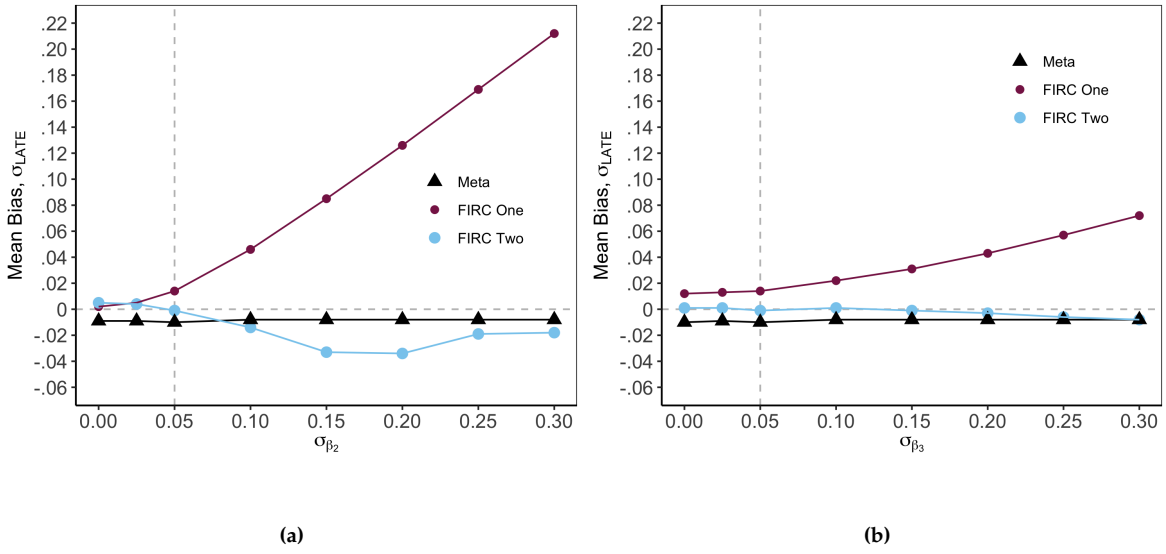


Figure 2.4: *The mean bias in the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), fixed intercepts random coefficients with random running variable coefficients (FIRC Two), and random effects meta-analysis (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*

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 2.5). 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 2.6). 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.

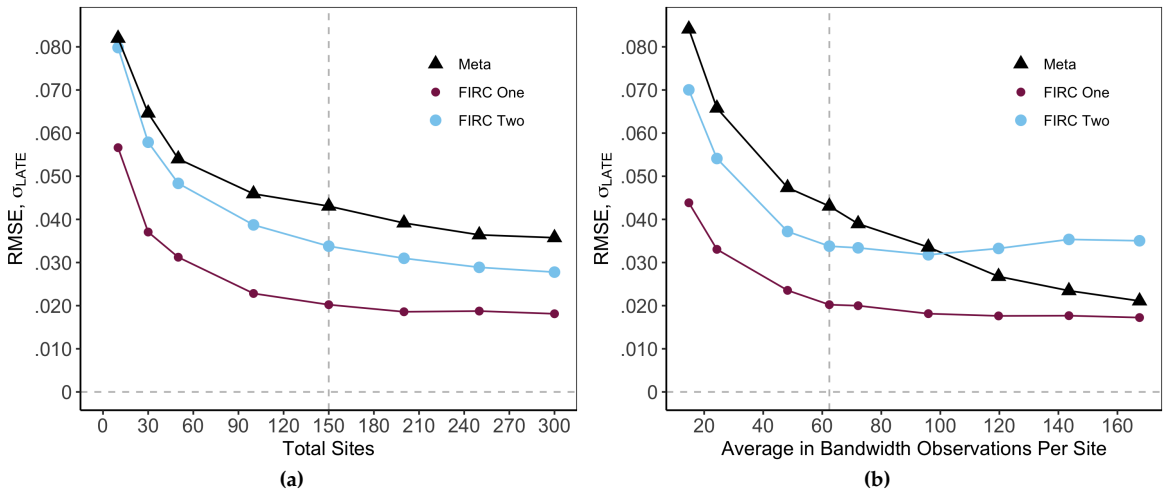
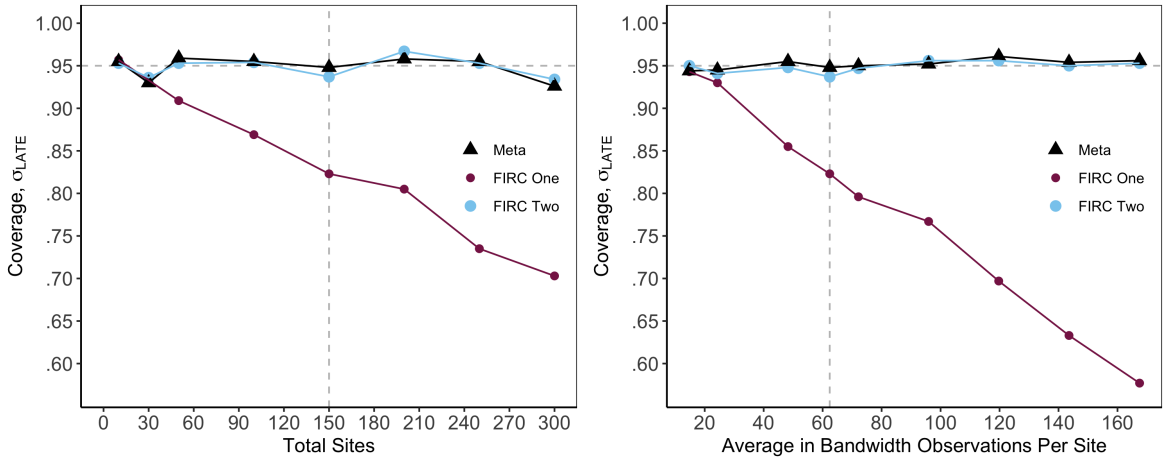


Figure 2.5: *The root mean squared error in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), and fixed intercepts random coefficients with random running variable 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 2.6: *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).*

2.4.3 The Meta Model

The Meta model works best out of the three models if there is cross-site variation in the running variable coefficients and the average number of in bandwidth observations is greater than 100. Once there are approximately 100 in bandwidth observations per site, the Meta model has less bias than both FIRC models (Figure 2.3b) and less error than the FIRC Two model (Figure 2.5b). The FIRC models partially pool data across sites, so they trade off additional modeling assumptions in exchange for increased precision in each site. When the average number of in bandwidth observations per site reaches 100, this trade off is no longer valuable, and the Meta model estimate has less bias and less error than the FIRC Two model.

2.4.4 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 variance in the running variable coefficients is detected, researchers should use the FIRC Two model unless the average in bandwidth observations per site is over 100, and then the Meta model should be used.

2.5 Massachusetts Education Proficiency Plan Example

2.5.1 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.

2.5.2 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 2.7). 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.

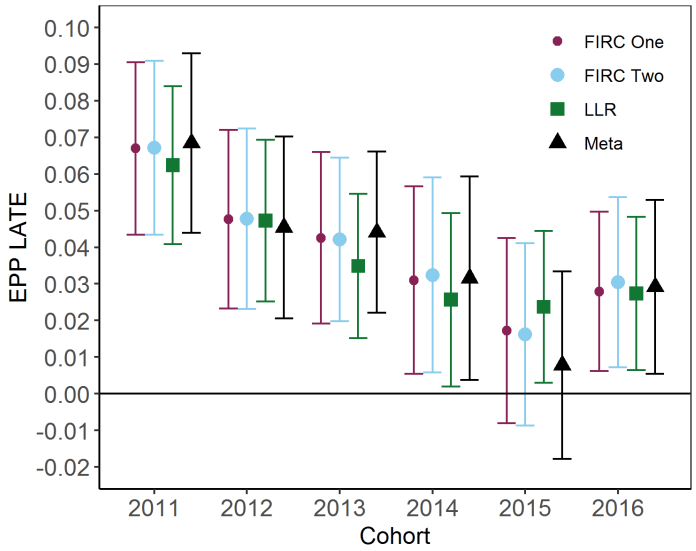
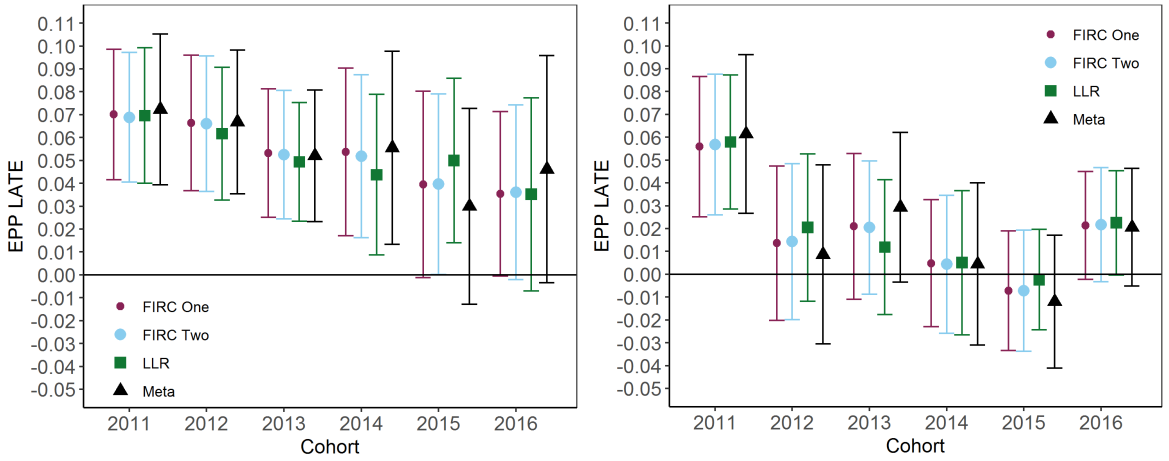


Figure 2.7: *The local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts.*

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 2.8). 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 2.8: *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 100, but the average number of in bandwidth observations per site was 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 2.9) 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.

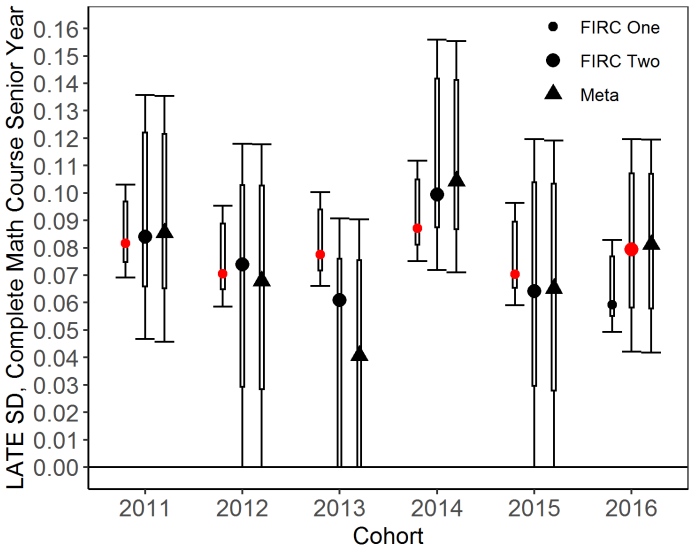
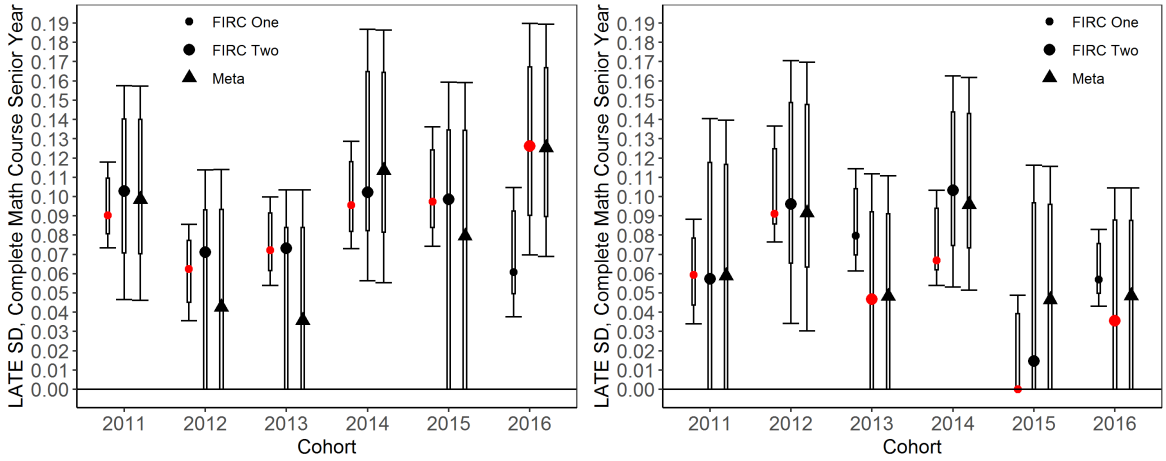


Figure 2.9: *The cross-high school standard deviation of the local average treatment effect of the math Education Proficiency Plans on the probability of completing a math course senior year across cohorts. The 95% and 80% confidence interval is marked for each point. In each cohort, the estimate marked in red is the preferred model.*



(a) *High Schools Requiring $\leq$ 4 Years of Math to Graduate*

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

Figure 2.10: *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 2.8. However, this type of heterogeneity analysis only captures the “systematic component" of treatment effect variation (Ding, Feller, and 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 2.10). 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 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 that did require four years of math to graduate, it is, however, 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.

2.6 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 (Bloom et al., 2017; Raudenbush et al., 2012; Reardon and Raudenbush, 2013; Reardon, Unlu, Zhu, and Bloom, 2014). 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 may 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 unrestricted FIRC model should be used when the average number of in bandwidth observations per site is less than 100, and the Meta model should be used when the average number of in bandwidth observations per site is greater than 100.

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.

Chapter 3

Local Randomization Regression Discontinuity Designs when Test Scores are the Running Variable

Abstract

Explanations of the internal validity of regression discontinuity designs (RDD) generally appeal to the idea that RDDs are “as good as" random near the treatment cut point. Cattaneo, Frandsen, and Titiunik (2015) are the first to take this justification to its full conclusion and propose estimating the RDD local average treatment effect (LATE) the same as one would a randomized experiment. This paper explores the implications of analyzing an RDD as a local random experiment when the running variable is a test score. I derive a formula for the bias in the LATE estimate estimated using the local randomization method, $a\rho\Delta$. Where a is the relationship between latent proficiency and the potential outcome absent treatment, ρ is the test reliability, and Δ is the distance between the treatment and control running variable value. I argue that this bias will make local randomization problematic for most test score RDDs and other RDDs that use human developed measures (e.g., medical tests).

3.1 Introduction

Regression discontinuity inference is sometimes explained as follows: two different students sit in two different classrooms taking the same math test. The two different students have the same math proficiency, but outside the window of one student's classroom, two dogs start barking at each other. The student with the dogs outside is distracted, misses a question they would normally get correct, and fails the math exam. The student with no dogs gets the question right and passes the math exam. The student who fails the exam is given extra math support, and the student who passes is not. Since the appearance of the barking dogs is random, the student who passes can be used as a control for the student who fails.

There is a problem with this story; on average, students who get the minimum passing test score are more proficient than students who get one point below passing. The barking dogs story looks at two individual students with identical proficiency, whose only difference is the presence of barking dogs. However, on average, students who get the minimum number of questions correct to pass will experience as many barking dogs as the students who fall one question short of passing. Therefore in expectation, students who fail the exam by one question are not a good control for students who pass the exam by one question because, on average, the passing students are one question more proficient than the failing students.

In practice, this limitation of the barking dog story is not a problem for most regression discontinuity design (RDD) research. Heuristic explanations of RDD in the barking dog genre are meant to provide the intuition for RDD. They are not intended to be formal articulations of the identification assumption underpinning RDD. Instead, most RDD research relies on the continuity assumption first articulated by Hahn, Todd, and Van der Klaauw (2001), which is that the potential outcome of interest is continuous through the cut point. This continuity assumption is most commonly operationalized using local linear regression (LLR). Under LLR, the outcome of interest (e.g., passing a math course) is modeled as a linear combination of the running variable (e.g., math test score) and treatment status (e.g., receiving extra math support). The treatment effect is solved for using a

traditional OLS estimator but only using points within some optimal bandwidth of the treatment cut point.

Still, there is a tension in the literature between relying on the looser continuity assumption for identification and a desire to treat RDD identification “as good as random" near the cutoff. Lee and Lemieux (2010) argue that as long as there is some measurement error in the running variable and individuals cannot precisely control their running variable value, then an RDD is equivalent to a randomized experiment in the region of the cutoff. Papers that provide overviews of RDD methods also frequently explain RDD as a local randomized experiment (Imbens and Lemieux, 2008; Lee and Lemieux, 2010). There is also empirical evidence that RDDs provide the same estimates of the local average treatment effect as RCTs (Chaplin, Cook, Zurovac, Coopersmith, Finucane, Vollmer, and Morris, 2018).

However, even researchers that appeal to an “as good as random" articulation of RDD analyze their data using the LLR method suggested by Hahn et al. (2001). Cattaneo, Frandsen, and Titiunik (2015) are the first to advocate analyzing RDDs as a locally randomized experiment. In this framework, as in a fully randomized experiment, the LATE is calculated by comparing the mean outcome of treated observations near the cut point to non-treated observations near the cut-point. This analysis method is appealing because it moves RDD analysis from model based inference to design based inference. It also allows standard errors to be estimated using randomization inference, which may more accurately capture the sample size used to estimate the LATE.

In this paper, I explore the implications of using the local randomization method for estimating RDD local average treatment effects when the running variable is a test score. Test scores are a common running variable choice in education becuase educational interventions are frequently assigned based on test scores. More broadly, test scores serve as a framework for examining the larger class of RDDs, where the running variable is a noisy measure of an underlying latent quantity.

I show there is bias in the local randomization RDD LATE estimator and that bias is a multiplicative combination of the relationship between the latent construct and the outcome,

the test reliability, and the distance between points on the running variable scale. This bias emerges because test scores are discrete, and the RDD LATE can not be estimated at the limit. This problem is not unique to test scores. All empirical data is to some extent discrete because all measurement methods have finite precision. In addition, sample size constraints require the pooling of data across different points on a continuous scale for it to be analyzed as a random experiment. I also show how this bias can escape detection in covariate based placebo tests and that matching on observable characteristics is slow to reduce the estimator's bias. Taken together, this paper provides a basis for assessing the conditions under which local randomization is an appropriate framework for RDD analysis and when this framework will produce unacceptably biased results.

3.2 Literature Review

LLR RDD analysis is a form of model based inference. LLR requires researchers to estimate the local average treatment effect by modeling the treatment and control outcomes at the cut point instead of observing them directly. RDD is especially model based when the running variable is discrete, and there is no possibility of observing both treatment and control observations directly at the cut point. Lee and Card (2008) argue that when the running variable is discrete, specification error in the modeling can lead confidence intervals derived from the Eicker-Huber-White (EHW) heteroskedasticity-robust standard errors to have insufficient coverage. Lee and Card (2008) proposed that when the running variable is discrete, confidence intervals should use standard errors clustered at the discrete values of the running variable.

However, Kolesár and Rothe (2018) demonstrate that standard errors clustered at the running variable do not successfully account for model misspecification error and have worse coverage than the EHW confidence intervals. They find that EHW confidence intervals have good coverage properties for RDD models with small and medium bandwidths. Kolesár and Rothe (2018) also provide two methods for estimating confidence intervals when the model bandwidth large or when the running variable only takes on a small number of

discrete values. In addition, other studies have developed methods for correcting confidence intervals in cases where large confidence intervals may cause concern (Armstrong and Kolesár, 2020; Calonico, Cattaneo, and Titiunik, 2014; Calonico, Cattaneo, and Farrell, 2018).

Local randomization analysis as proposed by Cattaneo et al. (2015) provides are an alternate solution to the original problem articulated by Lee and Card (2008). Treating the RDD as a randomized experiment allows for confidence intervals to be estimated using randomization inference. Randomization inference also has the added benefit of being a form of finite sample inference, instead of relying on large sample approximations (Cattaneo et al., 2015).

3.3 Latent Variable Regression Discontinuity Model

Under the potential outcomes framework the regression discontinuity model for an individual i can be written as follow:

$$Y_{0i} = g_0(X_i) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$$

$$Y_{1i} = Y_{0i} + g_1(X_i)$$

$$Y_i = \begin{cases} Y_{0i} & X_i \geq c \\ Y_{1i} & X_i < c \end{cases}$$

where Y_{0i} is the outcome absent treatment, Y_{1i} is the outcome with treatment, X_i is the value of the running variable, and c is cut value that determines treatment. The treatment effect $(Y_{1i} - Y_{0i})$ for individual i is therefore $g_1(X_i)$.

When X is a test score, X is a noisy estimate of a true latent proficiency θ. Under a classical test theory model, we can write X_i as a linear combination of θ and a random error term ω such that:

$$X_i = \theta_i + \omega_i,\ \omega_i \overset{iid}{\sim} N(0, \sigma_\omega^2)$$

Under a latent variable model framework, Y is not a function of X, the noisy estimate, but θ the underlying latent quantity X measures. Combining the classical test theory model with the potential outcomes RDD model, the RDD problem can now be restated as:

$$Y_{0i} = g_0(\theta_i) + \epsilon_i$$

$$Y_{1i} = Y_{0i} + g_1(\theta_i)$$

$$Y_i = \begin{cases} Y_{0i} & X_i \geq c \\ Y_{1i} & X_i < c \end{cases}$$

$$X_i = \theta_i + \omega_i$$

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2),\ \omega_i \overset{iid}{\sim} N(0, \sigma_\omega^2),\ Cor(\epsilon_i, \omega_i) = 0$$

3.4 Local Randomization Bias

This section uses the latent variable RDD model from the previous section to derive the expected bias in the local randomization treatment effect estimate. In order to do this I make three simplifying assumptions, over the interval $(c, c - \Delta)$:

1. $g_0(\theta_i)$ is linear.
2. The error term ω_i is homoskedastic.
3. $g_1(\theta_i)$ is constant and equals the treatment effect τ.

Simplifying assumptions one and two should not be considered overly restrictive. It is expected that sufficiently close the treatment cut $g_0(\theta_i)$ can be approximated with a linear function. Similarly, tests are designed to have consistent measurement error across the test score scale, and observations at two adjacent points should not have substantially different amounts of error. Even in a case where far apart observations are binned, the homoskedastic assumption holds for many tests. For example, in the 2014-2015 Smarter Balanced tests, which report test reliability by decile, have reliabilities that don't deviate by

more than .1 from the modal reliability in all deciles between the 2nd and the 10th in all grades 3rd through 11th in both ELA and Math (Smarter Balanced Assessment Consortium, 2016). However, simplifying assumption three is a strong assumption because it assumes a constant treatment effect across values of the running variable. In Appendix B.1, I provide derivations where this assumption is relaxed, and I arrive at the same bias estimate. Using these assumptions, the simplified RDD model is written as follows:

$$Y_i = a\theta_i + \tau \mathbb{1}(X_i < c) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N[0, \sigma_y^2]$$

$$X_i = \theta_i + \omega_i,\ \omega_i \overset{iid}{\sim} N[0, \sigma_\omega^2]$$

Local randomization inference treats the RDD as a randomized experiment in the region near the cut point. In the discrete case, I define the control analysis sample as the observations where $X = c$ and the treated analysis sample as the observations where $X = c - \Delta$, where Δ is the distance between points on the discrete running variable scale. Using a design based difference in means estimator, the treatment effect is estimated as:

$$\hat{\tau} = E[Y_1|X = c - \Delta] - E[Y_0|X = c]$$

$$= E[a\theta + \tau + \epsilon|X = c - \Delta] - E[a\theta + \epsilon|X = c]$$

$$= aE[\theta|X = c - \Delta] + \tau - aE[\theta|X = c]$$

According to Kelley's formula (Kelly, 1947):

$$E[\theta|X = x] = \rho x + (1 - \rho)E[X]\ , where\ \rho = \frac{\sigma_\theta^2}{\sigma_X^2}$$

Therefore the estimator bias can be written as:

$$\hat{\tau} - \tau = a(\rho(c - \Delta) + (1 - \rho)E[X] - (\rho c + (1 - \rho)E[X]))$$

$$\hat{\tau} - \tau = -a\rho\Delta$$

The bias is a function of three things: the relation between the potential outcome absent

treatment and the latent proficiency (a), the test reliability (ρ), and the distance between adjacent points at the cutpoint (Δ). If any of these three quantities are zero, then the local randomization bias is zero. In the limit, where $a = 0$ or $\rho = 0$, we can consider the RDD to be a fully randomized experiment. This might occur if X is not an observed test score but a randomized lottery number. The lottery number has no relationship to either true proficiency (θ) or the potential outcome absent treatment (Y_0). The case where $\Delta -> 0$ is where the data is truly continuous, and the RDD analysis converges to the idealized RD where the LATE is estimated at the true limit.

However, test score RDDs are not situations where this bias will generally be close to zero. Test scores are strongly correlated with a wide range of outcomes (Hanushek, 2009; Heckman, Stixrud, and Urzua, 2006), leading to large a values. Tests are designed to reliable, and conventional test reliabilities are between .8 and .95 (Ho and Reardon, 2015). Finally, test scores typically have a non-trivial value for Δ; for example, both of the 2019 SAT sections had a gap of about .1 standard deviations between score bins (College Board, 2019). Taken together, these facts mean that in most cases, local randomization is not going to be a good method for estimating the RDD LATE in studies that use test scores as a running variable.

3.5 Placebo Test

Cattaneo et al. (2015) partially address the potential for bias in their local randomization analysis framework. They explain that local randomization is only a valid framework for RDD analysis when there is a region around the cut point where there is no relationship between the running variable and Y_0. However, their method for testing this assumption still allows for situations where the RDD LATE will be biased.

Cattaneo et al. (2015) propose using a placebo test to determine a region around the cut point where this orthogonality assumption holds. In this test, Y is replaced with a covariate Z (e.g., student race), which is not be affected by the treatment. The difference in means test statistic is then estimated for increasingly large values of Δ to find the largest value such that the hypothesis that the new $\hat{\tau}_Z$ is zero can not be rejected. This value of Δ is then used

as the region where the orthogonality assumption holds.

To understand how this placebo test does not keep the local randomization RDD LATE estimate from being unbiased, I use a new version of the simplified RDD model where Z is the outcome instead of Y. The new model is as follows:

$$Z = b\theta_i + \tau \mathbb{1}(X_i < c) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N[0, \sigma_z^2]$$

By construction $\tau_Z = 0$ and so from the results in the previous section it can be concluded that:

$$\hat{\tau}_Z = -b\rho\Delta$$

The conditions of the placebo test imply that this $\hat{\tau}_Z$ will be just below the minimum detectable value ($MDV - \delta$). $\hat{\tau}_Z$ will end up close to the MDV because the placebo test procedure for determining Δ keeps increasing Δ until $\hat{\tau}$ is just below the MDV. Solving for $\rho\Delta$ leaves $\rho\Delta = -\frac{MDV-\delta}{b}$. In turn solving for the RDD LATE estimate bias, leaves a bias of $-\frac{a}{b}(MDV - \delta)$.

If the true proficiency (θ) is more predictive of the covariate (Z) than the outcome (Y), then the bias will be less than the MDV. In this case, when the true LATE is zero, the local randomization estimator will, in most cases, not detect a statistically significant effect even with nonzero bias. However, in many cases, θ will be more predictive of Y than Z. Test scores are specifically designed to be predictive of outcomes of interest to education policymakers, which are also the outcomes frequently measured in education policy research. Tests like the ACT and the SAT partially derive their validity from their ability to predict the GPA's of first-year college students. On the other hand, test makers try to minimize the amount to which they are just capturing immutable characteristics, which are precisely the Z variables commonly used in placebo tests.

3.6 Matching

Another potential method for dealing with the bias in the local randomization method is to match observations within score buckets. In this section, I will derive a new bias equation that accounts for matching and show that matching does a poor job of reducing the bias in the LATE.

Take a new example where Y still depends on θ, and X is a noisy estimate of θ. Now assume θ is a linear combination of an observed variable V and an unobserved residual U. Such that:

$$Y = a\theta_i + \tau \mathbb{1}(X_i < c) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N[0, \sigma_y^2]$$

$$X_i = \theta_i + \omega_i,\ \omega_i \overset{iid}{\sim} N[0, \sigma_\omega^2]$$

$$\theta_i = dV_i + U_i$$

Note that V is measured without error and that Y only depends on V through θ. The new matching estimate for the treatment effect is:

$$\hat{\tau} = \sum_V P(V = v)(E[Y_1|X = c - \Delta, V = v] - E[Y_0|X = c, V = v])$$

$$= \sum_V P(V = v)(a(E[\theta|X = c - \Delta, V = v] - E[\theta|X = c, V = v]) + \tau)$$

$$E[\theta|X = x, V = v] = \frac{\sigma^2_{\theta|V=v}}{\sigma^2_{X|V=v}} x + \frac{\sigma^2_{\omega|V=v}}{\sigma^2_{X|V=v}} E[X|V = v]$$

Define R^2 to be the percent of the variance of θ explained by V, such that $R^2 = \frac{\sigma^2_{dV}}{\sigma^2_\theta}$. Therefore:

$$Var(\theta|V = v) = Var(dV|V = v) + Var(U|V = v) = Var(U) = (1 - R^2)\sigma_\theta^2$$

$$Var(X|V = v) = Var(\theta|V = v) + Var(\omega|V = v) = (1 - R^2)\sigma_\theta^2 + \sigma_\omega^2$$

$$E[\theta|X = x, V = v] = \frac{(1 - R^2)\sigma_\theta^2}{(1 - R^2)\sigma_\theta^2 + \sigma_\omega^2} x + \frac{\sigma_\omega^2}{(1 - R^2)\sigma_\theta^2 + \sigma_\omega^2} E[X|V = v]$$

$$\hat{\tau} = \sum_V P(V = v)(\tau - a\Delta \frac{(1 - R^2)\sigma_\theta^2}{(1 - R^2)\sigma_\theta^2 + \sigma_\omega^2})$$

The new equation for the bias is:

$$\hat{\tau} - \tau = -a\Delta \frac{(1-R^2)\sigma_\theta^2}{(1-R^2)\sigma_\theta^2 + \sigma_\omega^2}$$

if we assume that X is standardized such that $\sigma_X^2 = 1$ then the bias reduces to:

$$-\hat{\tau} - \tau = -\frac{1-R^2}{1-R^2\rho} a\Delta\rho$$

This means that in the matching case, the bias is reduced by a factor of $\frac{1-R^2}{1-R^2\rho}$, which means that increasing R^2 reduces the bias slowly. The derivative of this scaling factor in terms of R^2 is $\frac{\rho-1}{(1-R^2\rho)^2}$. To provide some numerical intuition, if we take a case where $\rho = .9$ and R^2 is zero; the magnitude of the bias is the same as in the example with no matching. This is intuitive because if the matching variables have no explanatory power, they should not increase the accuracy of the treatment effect estimate. When $R^2 = .5$ the bias is 90% of the no matching bias, when $R^2 = .7$ the bias is 81% of the no matching bias, and when $R^2 = .9$ the bias is still 53% of the no matching bias. This suggests that matching is an inefficient means of reducing the local randomization bias.

3.7 Conclusion

Substantial effort goes into making most test scores accurate measures of latent proficiency. This effort runs at cross-purposes to creating a randomized experiment. In RDDs analyzed using local linear regression, this is not a problem. However, when the heuristics used to explain the intuition behind RDD are taken literally, then the gap between heavily designed test scores and a truly randomized experiment can create problems.

In this paper, I quantify the bias associated with estimating the RDD LATE as a local random experiment. This quantification is not intended to adjust the LATE estimates ex-post. Such adjustment would move the estimate back into the model based inference and would not be superior to a typical local randomization analysis. Instead, this paper is designed to clarify under what conditions a local randomization model is appropriate for estimating the RDD LATE. Test scores, which have high reliabilities and good predictive power, are unlikely

to make for RDDs that can be analyzed with local randomization. Similarly, medical tests, which like academic proficiency tests, are designed to be precise and predictive, are unlikely to be good candidates for local randomization. On the other hand, RDDs, where the running variable is not a measure heavily designed by experts (e.g., birthday, election totals), may make for more plausible usages of the local randomization method.

Conclusion

Regression discontinuity design was conceived by two psychologists, Thistlewait and Campbell, in 1960 (Card, Mas, and Rothstein, 2008; Thistlethwaite and Campbell, 1960). Over the next few decades, the method was periodically reinvented by different academic disciplines but mostly remained in obscurity until it was revived by empirical labor economists in the late 1990's (Angrist and Lavy, 1999; Black, 1999; Card et al., 2008; Van der Klaauw et al., 1997). Today regression discontinuity is a method that spans disciplines, and innovations in regression discontinuity methods are made across statistics, economics, political science, and psychology (Butler and Butler, 2006; Card et al., 2008; Cook, 2008; Ludwig and Miller, 2007; Schochet, 2009; Skovron and Titiunik, 2015).

Interdisciplinary methodological research keeps methodological insights from one academic field that may be useful in other academic fields from lying dormant for thirty years. This dissertation is a testament to the value of taking an interdisciplinary perspective to quantitative methods. In the first essay, I apply a psychometric model and simulation to research problems in economics. In the process, I develop a new psychometrics based bunching estimator, an estimator primarily found in economics. In the second essay, I pull from meta-analysis and a cross-discipline literature on impact variation to develop methods for estimating the cross-site treatment effect variance in regression discontinuity designs. In the third essay, I apply models and statistical formulas from psychometrics to calculate bias in local randomization regression discontinuity design.

Finally, if there is one core insight of this dissertation is that statistical methods benefit from a detailed understanding of the objects being analyzed and the models being used.

Regression discontinuity uses local randomization as a heuristic justification for identification, but in practice, it is a form of model based inference. Accounting for the regression discontinuity model is essential when transferring experimental methods to a regression discontinuity context, and disregarding the model altogether can lead to bias. Test scores are complex statistical objects, and not properly accounting for their generation process can distort analysis. Understanding these details is essential to producing accurate methods.

Appendix A

Appendix to Chapter 2

A.1 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:

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

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

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

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 may seem like a reasonable compromise between the fully restricted FIRC One and the fully unrestricted FIRC Two model. However,

the FIRC Three model does not perform better than the FIRC Two model even when there isn't variation in the treatment running variable interaction coefficient. Figure A.1 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 61.

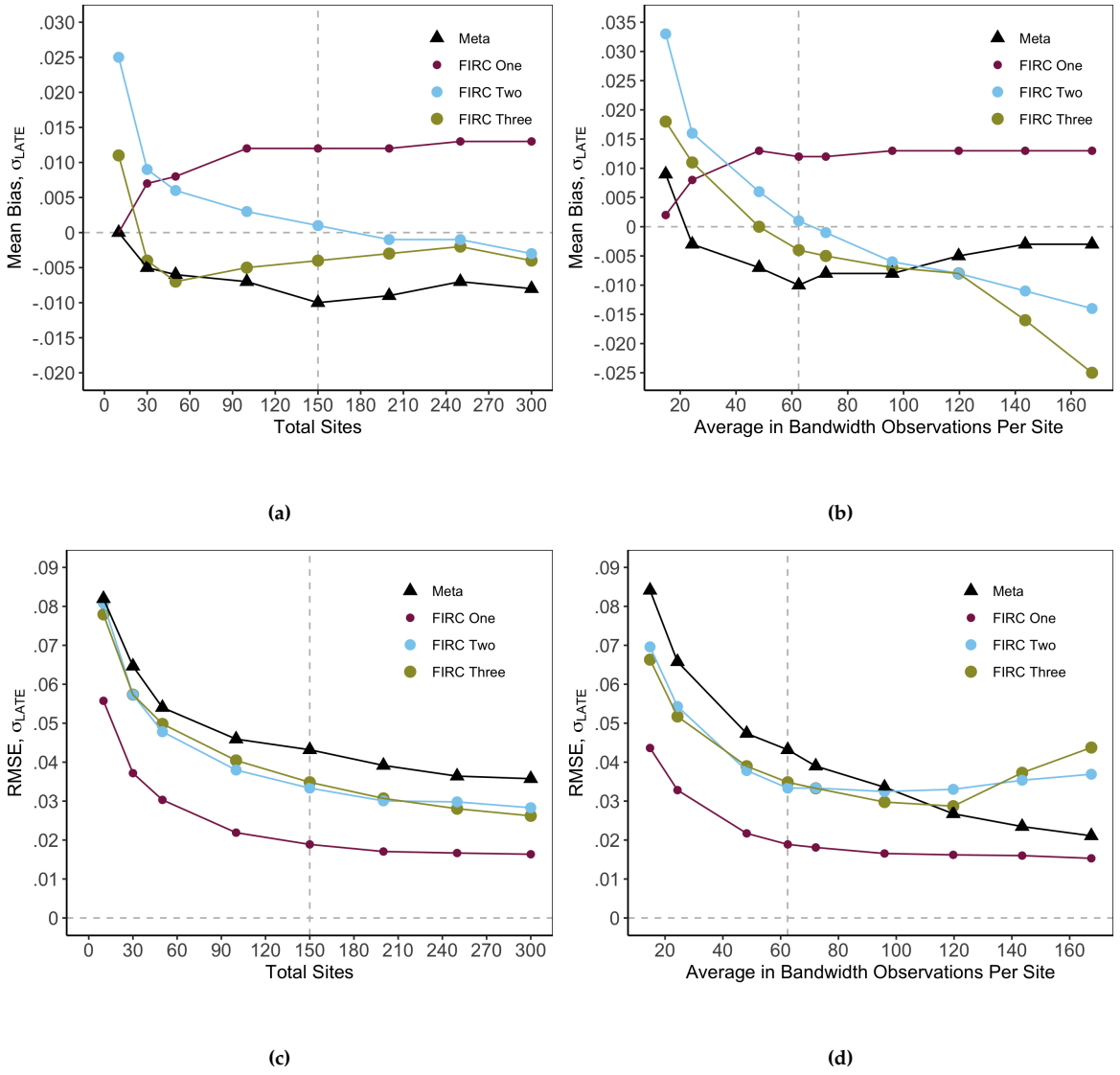


Figure A.1: *The mean bias (top) and root mean squared error (bottom) in the cross-site treatment standard deviation estimates across the random effects meta-analysis (Meta), fixed intercepts random coefficients with pooled running variable coefficients (FIRC One), fixed intercepts random coefficients 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).*

A.2 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 and Thompson (1971). This degrees of freedom correction is particularly important in a FIRC model because the site level fixed intercepts lead the model to have many coefficients that must be corrected for. However, none of the prior multi-site RDD studied that used a multi-level model to estimate cross-site treatment effect variance report whether they used REML or ML to fit their models McEachin et al. (2020); Raudenbush et al. (2012); Shapiro (2020). Figure A.2 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 A.3 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.

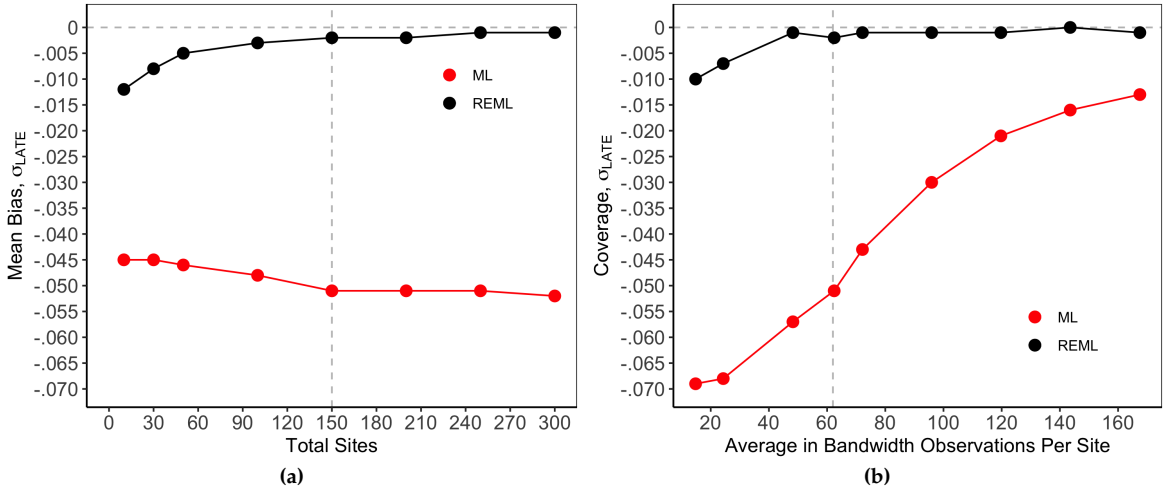


Figure A.2: *The mean bias in the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with pooled running variable coefficients (FIRC One) using Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation. In both panels the standard deviations 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 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).*

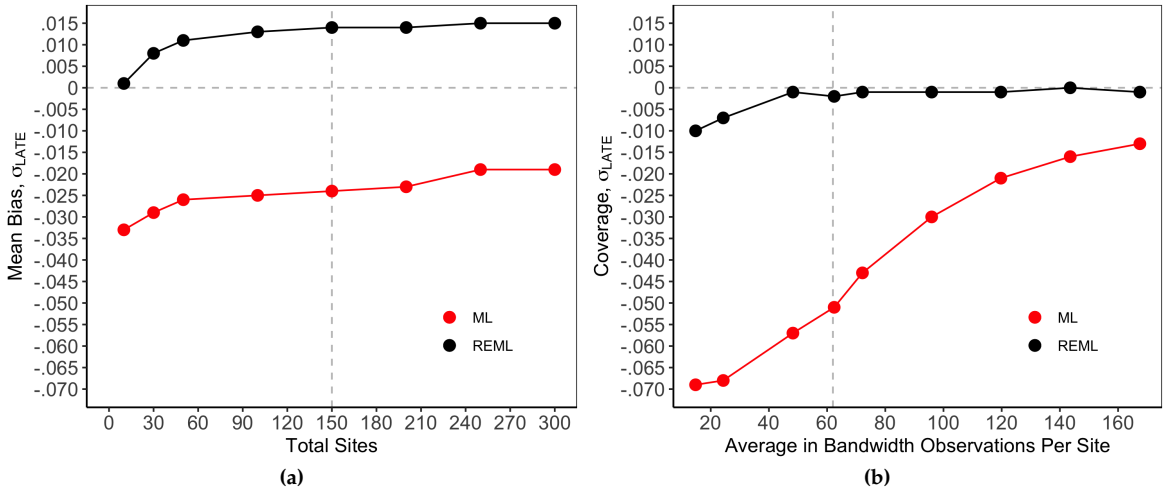
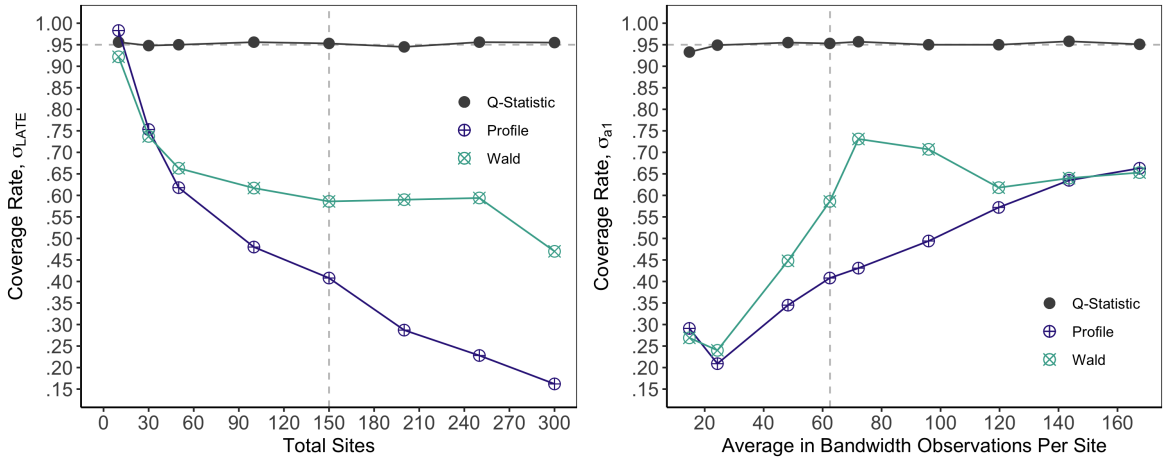


Figure A.3: *The mean bias in the cross-site treatment standard deviation estimates for the fixed intercepts random coefficients with random running variable coefficients (FIRC Two) using Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation. In both panels the standard deviations 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 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.3 Evaluation of Different Confidence Interval Methods for the FIRC Models

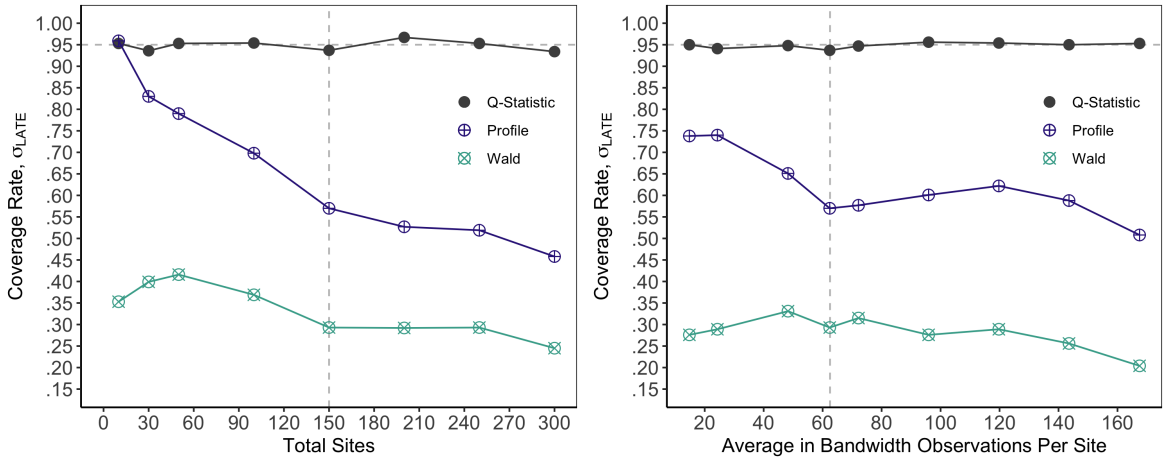
We tested three methods for estimating a confidence interval for the RDD FIRC cross-site treatment effect standard deviations estimates: Wald, Profiled, and Q-Statistic Inversion. In Figure A.4 and Figure A.5 we show the coverage rate for all three confidence interval types for the FIRC One and FIRC 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 A.2), 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 A.4: *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 A.5: *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 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).*

Appendix B

Appendix to Chapter 3

B.1 Derivation of Bias when the Relationship between the Outcome and Latent Proficiency Varies by Treatment Status

The assumption that $g_1(\theta_i)$ is constant and equals the treatment effect τ is dropped. Instead $g_1(\theta_i)$ is assumed to be a linear function $\beta + a_1\theta_i$. The new RDD model is written as follows:

$$Y_i = a_0\theta_i + a_1\mathbb{1}(X_i < c)\theta_i + \beta\mathbb{1}(X_i < c) + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N[0, \sigma_y^2]$$

$$X_i = \theta_i + \omega_i,\ \omega_i \overset{iid}{\sim} N[0, \sigma_\omega^2]$$

Under this model there is no longer a constant treatment effect and the LATE at $X = c - \Delta$ is therefore defined as:

$$\tau = E[Y_1|X = c - \Delta] - E[Y_0|X = c - \Delta]$$

$$= E[a_0\theta + a_1\theta + \beta + \epsilon_i|X = c - \Delta] - E[a_0\theta + \epsilon_i|X = c - \Delta]$$

$$\tau = E[a_1\theta + \beta|X = c - \Delta] = a_1E[\theta|X = c - \Delta] + \beta$$

The LATE is estimated as:

$$\hat{\tau} = E[Y_1|X = c - \Delta] - E[Y_0|X = c]$$

$$= E[(a_0 + a_1)\theta + \beta + \epsilon | X = c - \Delta] - E[a_0\theta + \epsilon | X = c]$$

$$= (a_0 + a_1)E[\theta | X = c - \Delta] + \beta - a_0 E[\theta | X = c]$$

The estimator bias can be written as:

$$\hat{\tau} - \tau = a_0 E[\theta | X = c - \Delta] - a_0 E[\theta | X = c] + a_1 E[\theta | X = c - \Delta] + \beta - a_1 E[\theta | X = c - \Delta] - \beta$$

$$= a_0 E[\theta | X = c - \Delta] - a_0 E[\theta | X = c]$$

According to Kelley's formula Kelley (1947):

$$E[\theta | X = x] = \rho x + (1 - \rho)E[X] \text{ , where } \rho = \frac{\sigma_\theta^2}{\sigma_X^2}$$

Therefore the estimator bias can be rewritten as:

$$\hat{\tau} - \tau = a_0(\rho(c - \Delta) + (1 - \rho)E[X] - \rho c - (1 - \rho)E[X])$$

$$\hat{\tau} - \tau = -a_0 \rho \Delta$$

References

AERA, APA, and NCME. *Standards for educational and psychological testing*. American Educational Research Association, 2014.

Mark Albanese. The testing column: Regrading essays and mpts-and other things that go bump in the night. *National Conference of Bar Examiners*, Mar 2016. URL https://thebarexaminer.org/article/march-2016/the-testing-column-regrading-essays-and-mpts-and-other-things-that-go-bump-in-the-night-2/.

American Psychological Association. Apa dictionary of psychology, 2021. URL https://dictionary.apa.org/psychometrics.

Joshua D Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics*, 114(2):533–575, 1999.

Joshua D Angrist, Parag A Pathak, and Christopher R Walters. Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, 5(4):1–27, 2013.

Timothy B Armstrong and Michal Kolesár. Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1):1–39, 2020.

Esteban M. Aucejo, T. Romano, and E. S. Taylor. Does evaluation change teacher effort and performance? quasi-experimental evidence from a policy of retesting students. *Review of Economics and Statistics*, pages 1–45, 2020.

Sandra E Black. Do better schools matter? parental valuation of elementary education. *The quarterly journal of economics*, 114(2):577–599, 1999.

Howard S Bloom, Stephen W Raudenbush, Michael J Weiss, and Kristin Porter. 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, 2017.

Eric T Bradlow and Howard Wainer. Some statistical and logical considerations when rescoring tests. *Statistica Sinica*, pages 713–728, 1998.

Simon M Burgess, Carol Propper, Helen Slater, and Deborah Wilson. Who wins and who loses from school accountability? the distribution of educational gain in english secondary schools. 2005.

Daniel M Butler and Matthew J Butler. Splitting the difference? causal inference and theories of split-party delegations. *Political Analysis*, pages 439–455, 2006.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

Sebastian Calonico, Matias D Cattaneo, and Max H Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.

David Card, Alexandre Mas, and Jesse Rothstein. Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218, 2008.

Matias D Cattaneo, Brigham R Frandsen, and Rocio Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, 3(1):1–24, 2015.

Duncan D Chaplin, Thomas D Cook, Jelena Zurovac, Jared S Coopersmith, Mariel M Finucane, Lauren N Vollmer, and Rebecca E Morris. The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2):403–429, 2018.

Ying Cheng and Cheng Liu. A short note on the relationship between pass rate and multiple attempts. *Journal of Educational Measurement*, 53(4):431–447, 2016.

BE Clauser and RJ Nungester. Classification accuracy for tests that allow retakes. *Academic Medicine*, 76(10):S108–S110, 2001.

College Board. 2019 SAT Suite of Assessments Annual Report, 2019.

Thomas D Cook. Waiting for life to arrive?: a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654, 2008.

Thomas S Dee, Will Dobbie, Brian A Jacob, and Jonah Rockoff. The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423, 2019.

Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

Karen M Douglas and Robert J Mislevy. Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, 35(3):280–306, 2010.

Frank J Dudek. The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, 86(2):335, 1979.

Kurt F Geisinger and Carina M McCormick. Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, 29(1):38–44, 2010.

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

Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

Eric A Hanushek. The economic value of education and cognitive skills. *Handbook of education policy research*, pages 39–56, 2009.

James J Heckman, Jora Stixrud, and Sergio Urzua. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24 (3):411–482, 2006.

Julian PT Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of

random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

Andrew D Ho and Sean F Reardon. Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, 40(2):158–189, 2015.

Huynh Huynh. Error rates in competency testing when test retaking is permitted. *Journal of Educational Statistics*, 15(1):39–52, 1990.

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

Truman Lee Kelley. *Fundamentals of statistics*. Harvard University Press, 1947.

Jeffrey R Kling, Jeffrey B Liebman, and Lawrence F Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119, 2007.

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304, 2018.

Daniel M Koretz. *Measuring up*. Harvard University Press, 2008.

David S Lee and David Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. Addison-Wesley, 1968.

Jens Ludwig and Douglas L Miller. Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122 (1):159–208, 2007.

Barbara Martinez and Tom McGinty. Students' regents test scores bulge at 65. *The Wall Street Journal*, Feb 2011. URL https://www.wsj.com/articles/SB10001424052748703445904576117793343465096.

Kinge Mbella, Min Zhu, Thakur Karkeend, and Hope Lung. The North Carolina Testing Program: Technical report 2012-2015: Mathematics assessments: End-of-grade 3-8 and end-of-course Math I. Technical report, North Carolina Department of Public Instruction, 2016. URL ttps://www.dpi.nc.gov/documents/accountability/testing/technotes/the-north-carolina-testing-program-technical-report-20122015-mathematics-assessments-end-of-grade-38-and-end-of-course-math-i.

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

Jason Millman. If at first you don't succeed setting passing scores when more than one attempt is permitted. *Educational Researcher*, 18(6):5–9, 1989.

Derek Neal and Diane Whitmore Schanzenbach. Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283, 2010.

New York State Education Department. Information Booklet for Scoring the Regents Examination in Integrated Algebra, 2009a. URL http://www.p12.nysed.gov/assessment/reports/2009/ia-tr-rv609.pdfl,. Accessed on November 23, 2016.

New York State Education Department. New York State Regents Examination in Integrated Algebra June 2009 Administration Technical Report on Reliability and Validity., 2009b. URL http://www.p12.nysed.gov/assessment/08-09memo/541ia-609.pdf,. Accessed on November 23, 2016.

North Carolina Department of Public Instruction. ABCs/AYP 2009 Accountability Report Background Packet, 2009. URL http://www.ncpublicschools.org/docs/accountability/reporting/abc/2008-09/backgroundpacket.pdf. Accessed on July 16, 2020.

H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

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

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

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

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

Randall Reback. Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5-6):1394–1415, 2008.

Peter Z Schochet. Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, 34(2):238–266, 2009.

Anna Shapiro. Over diagnosed or over looked? the effect of age at time of school entry on students receiving special education services, July 2020. URL http://www.edworkingpapers.com/ai20-259.

Christopher Skovron and Rocıo Titiunik. A practical guide to regression discontinuity designs in political science. *American Journal of Political Science*, 2015:1–36, 2015.

Smarter Balanced Assessment Consortium. Smarter Balanced Assessment Consortium: 2014-15 Technical Report, 2016. URL https://portal.smarterbalanced.org/library/en/2014-15-2015-16-interim-technical-report.pdf. Accessed on May 5, 2021.

Matthew G Springer. The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, 27(5):556–563, 2008.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

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

Wilbert Van der Klaauw et al. *A regression-discontinuity evaluation of the effect of financial*

aid offers on college enrollment. Number 97. New York University, Faculty of Arts and Science, Department of Economics, 1997.

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

Michael J Weiss, Howard S Bloom, Natalya Verbitsky-Savitz, Himani Gupta, Alma E Vigil, and Daniel N Cullinan. 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, 2017.

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