NCEE 2012-4025

U. S. DEPARTMENT OF EDUCATION

Replicating Experimental Impact Estimates Using a Regression Discontinuity Approach

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

Replicating Experimental Impact Estimates Using a Regression Discontinuity Approach

April 2012

Philip M. Gleason
Alexandra M. Resch
Jillian A. Berk
Mathematica Policy Research

Abstract

This report compares the estimated impacts of an education intervention based on an experimental design to the estimated impacts of the same intervention using a regression discontinuity (RD) design. The analysis uses data from two large-scale randomized controlled trials (RCTs) of education interventions—data from the IES restricted-use file from the Educational Technology Study and from the contractor-provided Teach for America Study analysis file. We found that the RD and experimental designs produced impact estimates that were not significantly different from one another, although the differences between the point estimates of impacts from the experimental and RD designs sometimes were nontrivial in size. We also found that manipulation of the assignment variable in RD designs can substantially influence RD impact estimates, particularly if manipulation is related to the outcome and occurs close to the assignment variable's cutoff value.

NCEE 2012-4025
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences (IES), under Contract ED-04-CO-0112/0006.

Disclaimer

The Institute of Education Sciences at the U.S. Department of Education contracted with Mathematica Policy Research to develop a report comparing estimates of a given education intervention based on a regression discontinuity design with those based on an experimental design. The views expressed in this report are those of the authors, and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
Rebecca A. Maynard
Commissioner

April 2012

This report is in the public domain. Although permission to reprint this publication is not necessary, the citation should be the following:

Gleason, Philip M., Alexandra M. Resch, and Jillian A. Berk. "Replicating Experimental Impact Estimates Using a Regression Discontinuity Approach." NCEE Reference Report 2012-4025. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2012.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

IES contracted with Mathematica Policy Research to develop the discussion of issues presented in this report. Philip Gleason, Alexandra Resch, and Jillian Berk are employees of Mathematica Policy Research. The authors and Mathematica do not have financial interests that could be affected by the content in this report. Data used in this study came from two studies previously conducted by Mathematica, the Evaluation of Education Technology, which was sponsored by IES, and the Teach for America (TFA) Study, sponsored by the Smith Richardson Foundation, Hewlett Foundation, and Carnegie Corporation.

Foreword

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts large-scale evaluations of education programs and practices supported by federal funds using study designs that support unbiased estimates of effectiveness; provides research-based technical assistance to educators and policymakers; and supports the synthesis and widespread dissemination of the results of research and evaluation throughout the United States.

In support of this mission, NCEE promotes methodological advancement in the field of education evaluation through analyses using existing data sets and explorations of new applications of technical methods. The results of these methodological investigations are published as commissioned, peer reviewed papers, under its Technical Methods Reports series, which is posted on the NCEE website at http://ies.ed.gov/ncee/pubs/. These reports are specifically designed for use by researchers, methodologists, and evaluation specialists. The reports address current methodological questions and offer guidance to resolving or advancing the application of high quality evaluation methods in varying education contexts.

This NCEE Technical Methods paper uses data from two large scale evaluations—the Education Technology and the Teach for America evaluations— to compare impact estimates based on regression discontinuity design (RD) methods with impact estimates based on the original experimental designs. The study finds that the RD and experimental designs produced impact estimates that were not significantly different from one another although the differences between the point estimates of impacts from the experimental and RD designs sometimes were nontrivial in size. We also found that manipulation of the assignment variable in RD designs can substantially influence RD impact estimates, particularly if manipulation is related to the outcome and occurs close to the assignment variable's cutoff value.

CONTENTS

TABLES

FIGURES

I. INTRODUCTION

A key challenge in the field of education research is to develop study designs that can generate unbiased and/or statistically consistent estimates of program or intervention impacts in cases in which an experimental design is not possible.[1] Regression discontinuity (RD) designs are becoming popular alternatives in such situations, as evidenced by the recently released What Works Clearinghouse (WWC) *Pilot Standards for Regression Discontinuity Designs* (Schochet et al. 2010). While RD designs have appealing theoretical properties (Hahn et al. 2001; Imbens and Lemieux 2008), questions remain about the performance of RD estimators in empirical applications. Existing simulation studies and within-study comparisons of RD and experimental estimates have yielded results consistent with theory and favorable to RD designs, but more needs to be known about the conditions under which the RD approach will perform well. This report compares the estimated impacts of each of two education interventions based on an experimental design to the estimated impacts of the same interventions using an RD design. By doing so, the report provides evidence on the performance of RD estimators in two specific contexts and, more generally, presents and implements a method for examining RD estimators that could be used in other contexts.

This study builds on an existing literature that attempts to replicate experimental findings using non-experimental methods. For more than twenty years, researchers have tested whether a comparison group design can produce reliable causal estimates. Many of the early replication studies were between-study comparisons that tested the sensitivity of a program's estimated impact to the choice of analytical methods and comparison groups. In an influential early study, Lalonde (1986) attempted to replicate impact estimates from the National Supported Work Demonstration using comparison groups constructed from Current Population Survey and Panel Study of Income Dynamics data, finding that the results were sensitive to the design and analytical methods used. More recent replication studies have emphasized within-study comparisons where researchers estimate a program's impact using random assignment and then estimate the same impact using some non-experimental technique based on data from the same study (Shadish 2000).

A. Overview of Approach

In the analysis presented in this report, we attempted to replicate the impact estimates from two recent experimental evaluations of education initiatives using an RD design. We used the actual data from these experimental studies and generated data sets structured so that they could have been generated under a well-implemented RD design of the same interventions. We then compared the resulting RD impact estimates to the experimental estimates based on the original data. After examining the performance of the RD design under these ideal conditions, we examined RD designs implemented under less than ideal conditions to investigate how the RD impact estimates changed when the assumptions under which the design was implemented were loosened to reflect the less optimal conditions that researchers often face.

We implemented this replication study using two experimental data sets from recent education evaluations. One of the experimental studies was the Educational Technology (Ed Tech) Study

[1] An impact estimator is consistent if any bias in its estimates goes to zero as the sample size goes to infinity. Thus, a consistent estimator may be biased in finite samples, but the bias becomes small when the sample size gets larger.

conducted by Mathematica Policy Research for the Institute for Education Sciences (IES) of the U.S. Department of Education (Dynarski et al. 2007). We obtained the restricted-use data set for this study from IES.[2] The other experimental study was the Teach for America (TFA) Study, conducted by Mathematica for the Smith Richardson Foundation, Hewlett Foundation, and Carnegie Corporation (Decker et al. 2004). We obtained the public-use dataset for this study from the Mathematica's Publications Coordinator.[3] The approach we use to compare impact estimates from RD and experimental designs could be used by other researchers with other data sources. This could provide additional evidence about the performance of RD designs for estimating impacts of education interventions in different contexts. In addition, the methods used in this report for aggregating evidence on the RD versus experimental estimates across the two different data sources examined here could be extended to incorporate evidence from other studies.

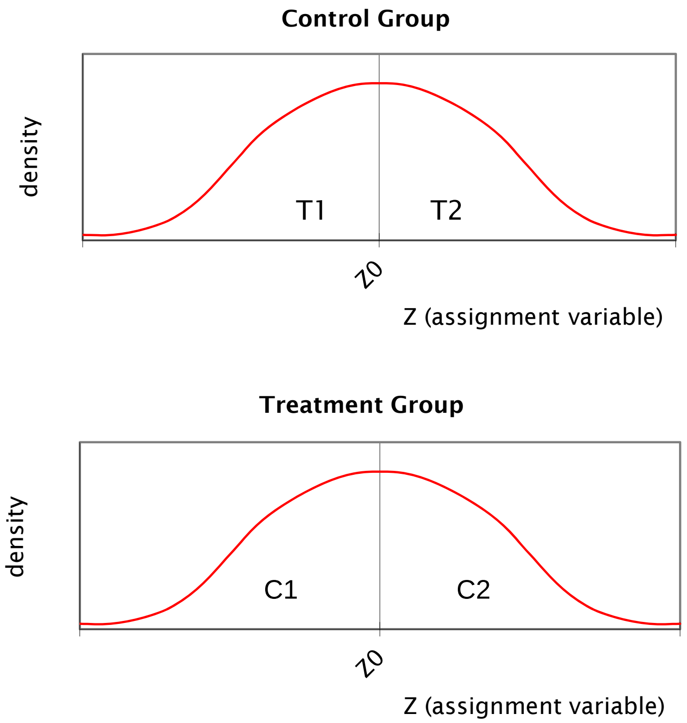
The centerpiece of the analysis for this replication study was the construction of RD analysis files, created by selectively dropping observations from the original experimental data files. These RD analysis files replicated the conditions under which an RD design might reasonably have been applied. Specifically, a baseline characteristic from the original study was used as the "assignment variable" (the variable in an RD design whose value determines sample members' treatment status), and a threshold, or cutoff, value for this assignment variable was selected. We then dropped all treatment group members with values of the assignment variable on one side of the threshold and all control group members on the other side. The resulting data set mimicked a situation in which treatment status was determined solely by the value of the assignment variable and an RD design would have been appropriate.

Figures I.1 and I.2 illustrate the process we used to construct the RD analysis file using a hypothetical example involving an assignment variable with a normal distribution and a cutoff value at its median. Figure I.1 shows the distribution of the assignment variable under the original experimental design for treatment and control group students separately. Because assignment to treatment status was random, the distribution of the assignment variable (a baseline characteristic) was independent of treatment status; that is, the distribution was identical for treatment and control group members. In the figure, the cutoff value Z_0 is shown at the median, so that the distribution is divided into two parts—T_1 and T_2 among the treatment group and C_1 and C_2 among the control group. The simulation of the RD data set from the original experimental data set involved the following three steps:

[2] Detail on the Ed Tech restricted-use file and instructions for obtaining a restricted use license can be found at http://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20074006.

[3] The public-use file can be obtained at no charge from Mathematica. Instructions can be found at http://www.mathematica-mpr.com/publications/howtoorder.asp.

Figure I.1. Distribution of Z in Experimental Study, by Treatment Status Split by Z_0



1. We selected the baseline characteristic to be used as the assignment variable (Z) along with the threshold value for that variable (Z_0). We used baseline test scores as the assignment variable and the median as the threshold value.

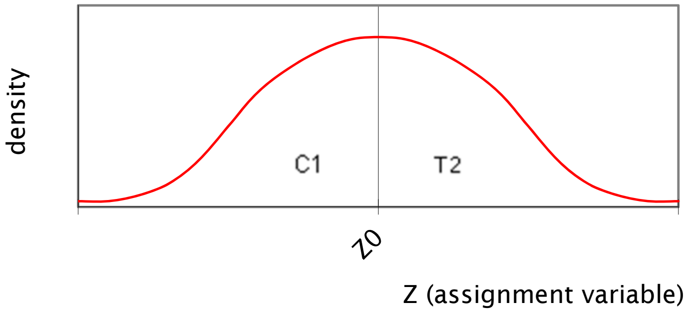
2. We divided the treatment group into the two subgroups T_1 and T_2, and control group into C_1 and C_2.

3. We dropped T_1 and C_2 so that the remaining data set included only T_2 and C_1.[4] The remaining data set had the same underlying distribution of Z as the original experimental data set, due to the random assignment design (Figure I.2). In other words, the RD sample as a whole represented the same population as the original experimental data set.

[4] In this example, sample members with high baseline test scores received the treatment and those with low scores did not. We refer to this scenario as RD High and to the resulting data as the RD High sample. In a separate analysis, we did the reverse by retaining treatment group members with baseline test scores below the median and control group members with baseline test scores above the median. We refer to this scenario as RD Low and to the resulting data as the RD Low sample.

Figure I.2. Distribution of Z in RD Data Set Constructed from Original Experimental Study Data Set

Constructed RD Data Set



The resulting data set matched one that would have resulted if treatment status had been assigned based on the value of the assignment variable. For example, a school district may have decided that a particular intervention was most appropriate for students in the top half of the achievement distribution and, thus, assigned students whose prior test scores were above the median to the intervention classrooms while assigning those whose test scores were below the median to control classrooms. Because treatment status in the original design was determined randomly, we could be confident that in this constructed data set, the baseline test score of sample members (Z) was the only baseline factor directly related to treatment status in the newly constructed RD data set. In other words, after controlling for baseline test scores, we were confident that none of the other baseline factors would be systematically associated with treatment status.[5]

With a data set constructed in this manner, we used sharp RD estimation methods to estimate impacts at the threshold point.[6] If one believes that the impact of the treatment is constant across the distribution of Z, this RD impact estimate should match the original RA impact estimate subject to sampling variability (which would arise because of the fact that a subset of observations from the original data set was dropped).

Three key limitations of the replication analysis presented in this report involve (1) its limited statistical power; (2) the artificial aspects of our method for creating the RD High and/or RD Low data sets; and (3) the fact that results of the replication exercise may be affected by idiosyncratic aspects of the two data sets examined. These limitations and our approach for addressing them are discussed below.

[5] Our RD analysis includes specification checks that would show whether such baseline factors were correlated with treatment status after controlling for baseline test scores.

[6] Sharp RD estimation methods are used when compliance to treatment status assignment is believed to be perfect or near perfect—that is, when all those assigned to the treatment group based on their assignment variable value actually receive the treatment while none of those assigned to the control group based on this value receive the treatment. By contrast, fuzzy RD methods may be used when there is some noncompliance to treatment status assignment. See Imbens and Lemieux (2008) for a discussion of sharp versus fuzzy RD estimation methods.

Statistical Power. The statistical power of the RD analysis conducted in this manner was lower than the statistical power of the original experimental analysis for two reasons. First, the replication of the experimental impact estimates using the RD approach relied on a data set that was roughly half as large as the original experimental data set. Second, the correlation between treatment status and baseline test scores in the RD analysis would substantially reduce the statistical power of the RD analysis relative to that of the experimental analysis, even if the sample size were the same, by a factor in the range of approximately 2.5 to 4 (Schochet 2008). The relative lack of statistical power of the RD impact estimates led, in turn, to relatively low statistical power of the comparison of the RD estimate of a particular impact parameter with the experimental impact estimate of the same parameter.

To address the limited statistical power of the replication exercise, we estimated two versions of the RD model—first with the RD High sample (shown in Figure I.2) and then with the RD Low sample. Similarly, as described above, we conducted this replication exercise using two separate experimental studies. Further, the TFA study had two major outcome variables (reading and math test scores), increasing the number of RD versus experimental comparisons we could make.[7] By aggregating all of these comparisons, we strengthened the basis for assessing the comparability of the RD approach and experimental approach to estimating impacts.

Artificial Creation of RD Samples. The fact that we used data from an actual experimental study to create an artificial data set that "could have" (but did not) come from an RD study has implications for what the replication exercise is actually testing. An important feature of this replication exercise is that it would result only from a situation in which conditions were optimal for the RD design. In the RD High (or RD Low) sample, the assignment variable was well defined and there was a clear cutoff value. In addition, we could be certain that there was perfect compliance with the assignment rule, as all students with baseline test scores below the cutoff were treatment group students who received the intervention while all those with scores above the cutoff were control group students who did not receive the intervention. In reality, however, researchers frequently face situations in which these optimal conditions do not hold. For example, there may be some uncertainty about whether sample members or program operators manipulated values of the assignment variable to either ensure that particular sample members received the intervention or did not receive it. Our basic replication exercise (presented in Chapters II through V of this report) does not allow for the possibility of this sort of manipulation, so comparing the RD estimates with the experimental estimates does not tell us anything about the possible role of manipulation in influencing the performance of RD designs to generate estimates of the impacts of education interventions.[8]

[7] Given that the reading and math scores of individual sample members are likely to be highly correlated, the boost in statistical power from adding a comparison between RD and experimental estimates for the second TFA test score is likely to be somewhat limited. We would have gotten a more substantial increase in power if the additional TFA outcome was independent of the initial test score outcome. As described in Chapter V, when we aggregated the various RD-experimental comparisons we conducted, we accounted for the non-independence of TFA reading and math scores in our tests of statistical significance.

[8] While the basic replication exercise does not address the issue of how manipulation of the assignment variable does or does not influence estimated impacts of education interventions based on RD designs, it does shed light on the ability of RD models to accurately model the relationship between the assignment variable and the outcome. Accurately modeling this relationship is key to the success of RD models in generating consistent impact estimates.

To explore the possible consequences of manipulation of the assignment variable, we conducted an additional analysis in which the process of creating the RD data set allowed for some manipulation of students' baseline test scores in order to alter their treatment status. This analysis, presented in Appendix C, sheds some light on the extent to which different types of manipulation influence RD impact estimates.

Idiosyncracies of the Ed Tech or TFA Studies. Any replication study that uses actual data from a real-world intervention to examine how one method for estimating impacts is similar to or differs from a second method is subject to the limitations of the data set being examined. The data reflect characteristics and outcomes for a single sample and capture the impacts of a single intervention. Idiosyncratic characteristics of the sample or of the intervention could affect the comparison of impact estimates produced by the two methods. In effect, the replication result reflects the performance of the RD estimator in the context of these two real world studies. RD designs may perform differently in other contexts.

In our case, we have results from the replication of experimental impact estimates using RD methods for two interventions and samples. More generally, the basic approach to replicating impact estimates from an experimental study presented in this report could be used in other situations to produce additional evidence on RD designs.

B. Prior Studies of the Performance of Regression Discontinuity Designs

This review of prior studies of RD designs focuses on studies that have attempted to assess the performance of RD estimation methods empirically. There is a large literature on RD estimation methods themselves, which is not discussed in this paper. See Cook (2008) for a history of the development of the RD design strategy and Imbens and Lemieux (2008) for a summary of current methodological practices and issues.

Trochim and Cappelleri (1992) conducted the first RD performance study of which we are aware. This study constructed a simulated data set that could have been generated by an RD design given a particular set of assumptions about the data-generating process (including the true impact of a given intervention). They found that in this idealized simulation setting, an RD model generated consistent estimates of a true treatment impact.[9] Cappelleri and Trochim (1994) built upon this work by testing a similar set of RD designs using experimental data from a clinical trial of a prescription drug and selectively dropping sample members to create an analysis file conducive to an RD design (an approach similar to what we do in this study). They found that the RD and experimental impact estimates were similar to one another. Though the article concludes by encouraging similar empirical studies, we are not aware of any other studies that used this approach to examine RD designs.

The other approach that has been used to assess the performance of RD designs involves generating both experimental and RD estimates of the same impact parameter using real data for both designs from the same study. Cook and Wong (2008) provided a critical review of the three such RD replication studies. Aiken et al. (1998) examined the effects of a college remedial writing

[9] We view this approach as being based on an idealized setting because all of the underlying relationships are known with certainty, and the implementation of the RD design is known (by construction) to follow perfectly the necessary conditions for using RD estimation techniques.

class at a large state university; Buddelmeyer and Skoufias (2003) examined data from PROGRESA, a Mexican conditional cash transfer program; and Black et al. (2007) estimated the effects of a reemployment services requirement among unemployment insurance recipients in Kentucky. In each case, particulars of the design made it possible to estimate the same (or a similar) treatment effect using both an experimental and RD design. The quality of implementation of the experimental and RD designs varied across studies, but Cook and Wong (2008) conclude that "…each study produced similar results across the experiment and regression-discontinuity study."

The within-study approach to assessing the performance of an RD design is useful, since it compares real-world applications of the RD method using actual data from random assignment studies. However, opportunities for conducting these studies are limited, and each may be influenced by idiosyncratic conditions of the study, characteristics of the data, or the quality of implementation of the RD and/or experimental design. Thus, these studies may not tell us much about whether an RD approach would be promising under different conditions than were present in those particular cases. A pure simulation study, on the other hand, would provide flexibility to examine the role of different aspects of the performance of RD estimators, but would not reflect the sort of real-world conditions that researchers are likely to confront when conducting a study. Thus, the approach we propose involves exploring the performance of RD designs using a combination of these strategies that relies on both real-world experimental data and a subset of these data artificially constructed to mimic conditions under which an RD design might typically be applied. By using data from actual experimental studies, real-world outcomes and idiosyncracies will be reflected. By artificially constructing a data set from the experimental data that might have been generated by an RD design, we could vary key design parameters to explore conditions under which RD results did or did not replicate experimental results.

Cook and Wong (2008) described the conditions that replication studies using within-study comparisons with a randomized experiment should attempt to meet. These conditions are listed below. While our replication effort was not truly a within-study exercise, since we artificially constructed the RD analysis file, the conditions are still relevant. In our design, we attempted to meet these conditions, to the extent possible.

1. "A within-study comparison has to demonstrate variation in the types of methods being contrasted—one comparison group has to be constructed via a random assignment mechanism and the other by whatever systematic mechanism is under test."
2. "The two assignment mechanisms cannot be correlated with other factors that are related to the study outcome." For example, outcomes for the sample members in the experiment should be measured using the same data source as outcomes for sample members in the non-experimental design.
3. "A quality within-study comparison also has to demonstrate that the randomized experiment deserves its status as a causal gold standard."
4. "It is also important that the non-experiment be a good example of its type." Cook and Wong (2008) go on to say that in an RD design it is important to appropriately handle the functional form issue, misallocations around the treatment status cutoff value in the assignment variable, and the lesser statistical power of the RD design.
5. "An experiment and non-experiment should estimate the same causal quantity." When comparing an impact estimate from an experiment with that from an RD design, this condition may be violated. We discuss this issue in Chapter II.

6. "A within-study comparison should be explicit about the criteria it uses for inferring correspondence between experimental and non-experimental results."

7. "The data analyst should perform the non-experimental analyses before learning the results of the experimental ones."

C. Preview of Remainder of Report

The remainder of the report provides details of the RD replication and results using the Ed Tech and TFA data. Chapter II provides a description of the RD and experimental estimation strategy we used to generate impact estimates under the two designs. The subsequent two chapters describe the two data sets used in the replication and present impact estimates based on the RD design, with the Ed Tech data and RD analysis presented in Chapter III and the TFA data and analysis presented in Chapter IV. Chapter V presents the experimental impact estimates from these two studies that served as a basis for assessing the performance of the RD estimator. That chapter also presents the comparisons of the experimental and RD estimates. Finally, Chapter VI provides a summary of the results.

II. STUDY DESIGN

In this chapter, we describe our approach to determining whether impact estimates based on a regression discontinuity (RD) model replicate those that would have come from an experimental study of the same intervention. In designing the study, we paid particular attention to several of the principles Cook and Wong (2008) provided for conducting good replication studies. They note, for example, that the non-experiment in the replication study should be a good example of its type. In the first section, we describe the steps we took to ensure that the regression discontinuity impact estimates resulted from current best practices. Cook and Wong (2008) also remind us that the experimental and non-experimental impact estimates should represent the same causal quantity. In comparing experimental and regression discontinuity models, this is not necessarily the case. The second section describes the specific experimental model and specifications we estimated and the steps we took to ensure that we were comparing estimates of the same impact parameter. Finally, it is important to explicitly define how the results of a replication effort will be judged prior to observing the impact estimates. In the third section of this chapter, we describe our standard for successful replication and we discuss some alternative approaches.

A. Estimating Impacts Using a Regression Discontinuity Design

In a regression discontinuity study, the assignment of students or other subjects to a given intervention is determined by the value of a predictor (or assignment variable) and on which side of a fixed threshold or cutoff value it falls. In our study, the assignment variable is the pretest score of students. For students with pretest scores very close to the cutoff, it may be reasonable to think of their treatment status as randomly assigned. Treatment students with scores just above the cutoff are likely to be very similar to comparison students with scores just below the cutoff. Unfortunately, researchers rarely have a sufficient amount of data to limit their analysis to students immediately surrounding the cutoff.

Instead, RD studies typically look beyond the data immediately surrounding the cutoff and use all of the data at their disposal, or at least all of the data within some fairly broad interval around the cutoff. In these RD designs, the performance of the estimator is dependent on properly modeling the relationship between the assignment variable and the outcome, since the assignment variable is—by definition—strongly associated with treatment status. This association is quite different from an experimental study where you would expect no correlation between student pretest scores and randomly determined treatment status. The association between the assignment variable and treatment status in an RD design is acceptable as long as the underlying relationship between the assignment variable and the outcome is smooth (aside from any treatment effect at the cutoff value) and the functional form of the relationship is properly modeled.

This section describes how we produced the main RD impact estimates to be compared with the RA estimates described in the next section.[10] The RD design included both a graphical analysis of the relationship among the assignment variable, treatment status, and outcome, as well as a more formal statistical estimate of program impacts. We also conducted a variety of specification tests to

[10] We performed the RD analysis before estimating impacts based on the experimental data, so that we would be blind to the target experimental estimate at the time we had to make decisions about the RD model specification.

examine the validity of the RD model. Our analysis conforms to the WWC *Pilot Standards for Regression Discontinuity Designs* (Schochet et al. 2010). The WWC standards include standards on the integrity of the assignment variable, attrition, continuity of the outcome-assignment variable relationship, and functional form and bandwidth.[11]

We estimated the impacts of the Ed Tech and TFA interventions using both parametric and nonparametric RD approaches. The basic parametric approach involved modeling the outcome test, or posttest, score as a function of the pretest score (the assignment variable) and treatment status. We estimated the following model:

$$(1)\ y_{ij} = \alpha_0 + \alpha_1 T_{ij}^{RD} + m(\delta, Z_{ij}) + \alpha_2 X_{ij} + \eta_j + \varepsilon_{ij}$$

where y_{ij} is the outcome test score of student i in classroom j, T_{ij} is a treatment indicator for the student, Z_{ij} is the pretest score, $m(\delta, Z_{ij})$ is a flexible function of the pretest score and a vector of parameters, X_{ij} is a vector of other baseline characteristics potentially influencing the outcome including indicators for the random assignment blocks and student-level covariates, η_j is a classroom-level error term, ε_{ij} is a student-level error term, and α_1 is our coefficient of interest—the impact of the intervention on the outcome. The Xs increase precision of the estimated treatment effect by consuming residual variance. We estimated this model using Generalized Least Squares (GLS) with a classroom-level random effect.

This model is quite similar to a basic experimental estimation equation, with the important difference that in the RD version, Z is not an "irrelevant" regressor but has a known relationship with treatment status. The simplest version of $m(\delta, Z_{ij})$ is $\gamma_1 Z_{ij}$, which imposes a linear relationship between the outcome and the pretest score that is the same on both sides of the cutoff. This can be generalized to allow for different linear slopes on each side of the cutoff or to allow for a nonlinear relationship. The parametric specifications we estimated potentially included linear, quadratic, and cubic terms, and allowed the function to differ on either side of the cutoff.

While graphical analysis provided the foundation for our parametric RD impact estimates, we used explicit procedures for selecting our optimal specification. We initially estimated a linear specification and then sequentially added higher order terms to the specification, testing their significance. To do so, we conducted a significance test to examine whether the higher-order terms

[11] To meet evidence standards without reservations, a study must meet each of the four individual standards without reservations. A study can meet evidence standards with reservations if the study meets the first, second, and fourth standards with or without reservations. The first standard requires statistical and institutional evidence that there was no systematic manipulation of the forcing variable. In this study, we constructed the RD, so we know there is no institutional possibility for manipulation, but we also present statistical evidence. The second standard on attrition requires that the study meet the WWC RCT standards for attrition. The third standard requires evidence that in the absence of the intervention, there would be a smooth relationship between the outcome and the assignment variable. In Chapters III and IV, we conduct appropriate sensitivity tests to examine whether key baseline covariates are continuous at the cutoff and whether there are discontinuities in the outcome-assignment variable relationship at values away from the cutoff. The final standard concerns the statistical modeling of the relationship between the assignment variable and the outcome. As recommended by the WWC, we have included extensive graphical analysis, allowed the estimated relationship to vary on either side of the cutoff, and used systematic approaches to select the most appropriate parametric and nonparametric specifications.

we had just added to the model (modeling separate relationships on either side of the cutoff) were jointly different from zero—that is, whether we could reject the null hypothesis that each of the coefficients on these terms was equal to zero. If we could not reject his null hypothesis, we dropped these higher order terms and selected the previous specification as the optimal parametric specification. If the higher order terms were significant, we added the next set of higher order terms. For example, if the quadratic terms were significant, we would add cubic terms. If the cubic terms were not jointly significantly different from zero, the quadratic model was our optimal parametric specification.

We used local linear regression, recommended by Imbens and Lemieux (2008), for estimating RD impacts nonparametrically. The simplest version of this procedure is equivalent to running linear regressions for a subset of data on each side of the cutoff.[12] For these local linear estimates, a key decision involves the choice of a bandwidth of data to be used in the estimation. As is common in the RD literature, we used a procedure for selecting an optimal bandwidth and then tested the sensitivity of these results for bandwidths half and twice that value. To select the optimal bandwidth, we used the Imbens and Kalyanaraman (IK) (2009) optimal bandwidth procedure. The IK method is a data-dependent method for choosing the bandwidth that is asymptotically optimal.[13]

In addition to estimating program impacts using the parametric and nonparametric RD methods described above, we used several specification tests to gauge whether our estimates were being driven by the failure of one or more of our assumptions about the continuity of underlying variables at the cutoff point. The RD design assumes that the only difference between observations just above and below the cutoff is their assignment to treatment. The validity of the RD method to determine impact estimates relies on the assumption that all baseline covariates are smooth through the cutoff point, making assignment to treatment "as good as random." These tests focus on ruling out discontinuities in the baseline and outcome variables that would create doubt that this important assumption holds. We focused on two categories of discontinuities: discontinuities at the cutoff value in the relationship between the pretest (assignment variable) and other baseline variables, and discontinuities in the assignment variable-outcome relationship at points other than the cutoff point.

[12] We used a rectangular kernel for the local linear estimates. We also considered a more sophisticated kernel function that would have weighted the observations differently based on their distance from the cutoff. However, Imbens and Lemieux (2008) point out that if the kernel choice changes the estimate, the estimate is also likely very sensitive to the bandwidth. Thus, we used only the rectangular kernel but followed their recommendation of using a rectangular kernel and testing for sensitivity to the bandwidth choice.

[13] Alternatively, we could have used a cross-validation procedure for selecting an optimal bandwidth, as described by Ludwig and Miller (2005). However, an advantage of the IK procedure over the cross-validation procedure is that the latter often requires the subjective judgment of the researcher for choosing among various potential optimal bandwidths. We compared both methods and found that the bandwidths selected by the IK algorithm were within the range of bandwidths suggested by the cross-validation procedure.

B. Estimating Impacts Using an Experimental Design

We estimated experimental impacts after the RD analysis was complete to ensure that our RD model selection was not influenced by knowledge of the experimental target. Although experimental impact estimates for EdTech and TFA had already been published at the time we estimated the RD models, our experimental models and analysis samples differed somewhat from those used in the original impact studies.

This main target impact estimate was based on the following experimental model:

$$(2) \quad y_{ij} = \beta_0 + \beta_1 T_{ij}^{RA} + \beta_2 Z_{ij} + \beta_3 X_{ij} + v_j + e_{ij}\,,$$

where y_{ij} is the outcome test score of student i in classroom j, T_{ij} is a treatment indicator for the student, Z_{ij} is the pretest score, X_{ij} is a vector of other baseline characteristics potentially influencing the outcome, v_j is a classroom-level error term (that is, a classroom random effect), and e_{ij} is a student-level error term.[14] In the model, we also controlled for the random assignment block. Indicators for these random assignment blocks are included in the vector X_{ij}. The estimate of coefficient β_1 is our experimental estimate of the impact of the intervention on student test scores. The pretest score (along with the other baseline characteristics) is, based on the experimental design, uncorrelated with treatment status so its inclusion should not affect the experimental estimate.

One important difference between the RD and experimental estimators is that the experimental estimator (β_1) represents the average treatment effect (ATE) for all students, while the RD estimator (a_1) represents the local average treatment effect (LATE), or the impact on students with test scores close to the cutoff value of the assignment variable. If impacts are constant across the range of values of the assignment variable, the ATE and LATE will be equivalent. However, if impacts are not constant across this distribution, the ATE and LATE will likely differ and so the treatment effects arising from the RD and experimental designs will likely differ as they will be providing estimates of different treatment effect parameters, even if both produce consistent estimates of these parameters.[15]

[14] Note that in the specification shown in equation (2) we used GLS to estimate a linear regression model that includes a random classroom effect along with the random student-level error term. The random classroom effect is intended to capture clustering of student outcomes among students within the same classroom. An alternative specification for modeling this sort of error structure, frequently preferred by education researchers, would be a hierarchical linear model (HLM) with student and classroom levels. This type of model was used in both the original Ed Tech and TFA experimental studies. We chose this different approach to specifying the model here so that the specification of the experimental model would be analogous to that of the RD model described above and any differences in the estimates would be less likely to be due to differences in the specification of the experimental and RD models. The model we estimate is quite similar to the corresponding HLM, estimated without imposing as many functional form assumptions. See section 24.6 of Cameron and Trivedi (2005) for a discussion of HLM and how it relates to other models for clustered data.

[15] It is theoretically possible that the LATE—evaluated at the cutoff of the assignment variables—could be equal to the ATE even if the treatment effect varies across values of the assignment variable. This could happen, for example, if the relationship between the assignment variable and treatment effect was linear and the cutoff value was at the mean of the assignment variable distribution.

A fair replication test requires that the experimental and non-experimental method estimate the same causal quantity. If our experimental estimate is the ATE and the RD estimate is the LATE, it is not necessarily reasonable to judge the performance of the RD method by comparing the two estimates. Instead, we estimated a local experimental impact using the restricted set of experimental data close to the cutoff score.[16] We use this local experimental estimator as a basis of comparison for the RD impact estimate (although we also estimated and present the full sample experimental impact estimate).

C. Comparing the RD and Experimental Impact Estimates

The final step in the replication exercise was to compare the RD and experimental estimates. In previous sections, we established procedures for choosing the primary RD and experimental estimates. While we estimated a variety of specifications, our overall judgment of the success of the replication effort was based on a comparison of these primary estimates. A variety of criteria could be used to assess the comparability of the RD and experimental estimates. Our primary criterion was whether the two estimates were statistically different from one another. Specifically, we estimated the difference between the RD and experimental estimates along with its standard error, and tested the null hypothesis that the difference was equal to zero versus the (two-sided) alternative hypothesis that this difference was not equal to zero.[17] To place the results of this significance test into broader context, we also present the point estimate and 95 percent confidence interval of the difference between the RD and experimental impact estimates.

To estimate the standard error of the difference between the RD and experimental impact estimates, we performed a bootstrap procedure. We first generated a set of 1,000 replicate samples from the original data by selecting a sample at random (and with replacement) from the original analysis file, with each replicate sample having the same sample size as the original analysis file. To account for the clustering of students within classrooms, we used a clustered bootstrap process. Instead of sampling individual students with replacement, the clustered bootstrap samples classrooms with replacement.[18] For each replicate sample, we: (1) estimated the experimental impact using the experimental model; (2) determined the optimal parametric specification and estimated the parametric RD impact;[19] (3) determined the optimal bandwidth and estimated the nonparametric RD impact; and (4) calculated the differences between each RD estimate and the experimental impact estimate.[20] After repeating the four-step process for 1,000 replicate samples, we calculated the standard deviation across replicate samples of the calculated difference between the RD and experimental estimates. This was our estimate of the standard error of the estimated difference between RD and experimental impacts based on the original sample, and was used to determine

[16] For each data set, we defined the restricted sample using an average of the optimal RD bandwidth from the RD High sample and the RD Low sample.

[17] This is the same criterion used by Black et al. (2007).

[18] This bootstrap process did not sample students within classrooms and therefore may understate the variance.

[19] For each of the 1000 bootstrap samples we determined the optimal specification for the RD estimate, so our standard error incorporates the variation that comes from the first stage choice of a specification.

[20] In practice, we estimated both RD High and an RD Low impact estimates (as well separate estimates of the difference between RD and experimental impacts for RD High and RD Low) for each replicate sample.

whether the two models produced impact estimates that were statistically different from one another.

The advantage of using the significance of the difference between the RD and experimental impact estimates as a basis for assessing the comparability of the two designs is that it is an objective criterion well grounded in statistical theory. Conclusions based on this test are less easily influenced by researchers' subjective opinions or unconscious biases. On the other hand, this significance test is best suited to determine whether there is sufficient evidence to conclude that two estimates are different from one another, rather than to determine whether the estimates are the same as one another.[21] In addition, the test is structured so that the lower the statistical power of the test, the more likely is a null finding and a conclusion that the RD and experimental designs produce impact estimates that are not significantly different from one another.

Due to these issues, we also used alternative criteria for assessing the correspondence between the RD and RA results. In particular, we compared the sign and general magnitude of the RD and experimental impact estimates, focusing on whether the general pattern of results across our two studies and the different test subjects was similar. We also examined whether the RD and experimental estimates had the same level of statistical significance.[22] Finally, we used our bootstrap replicate samples to compare the overall sampling distribution of RD and experimental impact estimates.

[21] An alternative approach would be the use of "equivalency testing," which is a statistical test of whether two estimates fall within the same range of values—that is, whether they are approximately the same (Rogers et al. 1993; Barker et al. 2002). In this type of test, the null and alternative hypotheses are the opposite of what they are in the significance test described above—the null hypothesis is that the two estimates differ by more than some predefined value and the alternative hypothesis is that the two estimates are within that predetermined value of one another. The downside of this approach is that it requires the researcher to select the amount by which the two estimates will be allowed to differ from one another to be considered equivalent, and this decision is often arbitrary.

[22] An important limitation of using the level of significance as a criterion for assessing the comparability of the RD and experimental designs is that in our primary analysis, the experimental design has substantially more statistical power than the RD design. Thus, even if the impact estimates were identical, their levels of statistical significance would not necessarily match.

III. RD ESTIMATION OF THE IMPACT OF ED TECH

This chapter describes the data from the Ed Tech study that we used in our replication effort, and presents the results from estimation of our basic RD specifications. We first briefly describe the Ed Tech study and the associated data and then discuss our RD estimates using the RD High sample. Key estimates from analysis of the RD Low and RD High samples are presented at the end of the chapter, and the details of the RD Low analysis are included in Appendix A.

A. Description of Study and Data

We used first year follow-up data from the Ed Tech study conducted by Mathematica for IES (Dynarski et al. 2007) to conduct this replication exercise.[23] The Ed Tech study was designed to examine whether the use of education technology in the classroom leads to higher achievement test scores for students, as called for by the No Child Left Behind Act of 2002. The study used a random assignment design to estimate the impact of reading and mathematics software products on test scores among elementary and secondary school children, covering four groups—grade 1 reading, grade 4 reading, grade 6 math, and high school algebra.

Promising software products were identified from public submissions of 160 products for consideration. The Ed Tech study team narrowed the pool of products to 40 based on previous evidence of effectiveness and the feasibility of implementing the product on a national scale. An outside panel evaluated these products and the U.S. Department of Education chose 16 for inclusion in the study. Districts that did not already use the study products were recruited to participate in the study, and teachers within participating schools volunteered to participate. The design required at least two participating teachers in each grade level within each school, but some schools had more than two. Within each school and grade, teachers were randomly assigned either to a treatment group, where they would implement one of the products in their classroom, or to a control group, where they did not receive the product.[24] When there was an even number of participating teachers in a school and grade, the teachers were evenly split between the treatment and control groups. In cases with an odd number of participating teachers, a larger number were randomly assigned to the treatment group. The final sample included 238 treatment teachers and 190 control teachers. Teachers in the treatment classrooms were trained to use one of the study products and implemented it in their classroom for the 2004-05 school year. Teachers in the control classrooms were not allowed to use the study products but could use computers and other technology as they normally would.

[23] See Dynarski et al. (2007) for more information on the Ed Tech study and its findings. A second year report (Campuzano et al. 2009) has been recently released, but we focus here only on first-year results.

[24] Students were not randomly assigned to classrooms, but the random assignment of the intervention to classrooms implies that the average baseline characteristics of students in the treatment classrooms were not systematically different from the average baseline characteristics of students in the control classrooms. Dynarski et al. (2007) found that differences between the two groups in each of the three grades we are using in this analysis were not statistically significant.

Students were tested at the beginning and end of the 2004-05 school year, the year in which the products were in place. Students in grades 1, 4, and 6 whose parents consented for them to participate in the study were tested using versions of the Stanford Achievement Test. The students in each grade were given comparable pre- and posttests. Algebra students were tested using an End-of-Course Algebra Assessment produced by ETS. Impact estimates were based on differences in end of year, or posttest, scores for the students in treatment classrooms compared to students in control classrooms. The first-year report was designed to evaluate the effectiveness of technology, not the effectiveness of individual products, so it presented average impacts across software products for each of the four subjects.[25]

Data from the Ed Tech study are appropriate for this replication exercise for several reasons. The study employed a relatively straightforward random assignment process, with the classrooms of eligible teachers from within participating schools randomly assigned into either the treatment or control group. This process resulted in a large sample of consenting students in these classrooms.[26] Overall, the Ed Tech evaluation sample consisted of 9,424 students in the classrooms of 428 teachers in 132 schools and 33 districts.[27] This included 5,399 students in the treatment group and 4,025 in the control group. Finally, the baseline test administered to the student sample provides a good candidate for the assignment variable in our RD analysis. One might imagine an intervention being provided only to the highest performing students, and thus performance on a standardized test might be used to determine students' eligibility for the intervention if random assignment were not possible.

We used a subset of the original Ed Tech data in the current study, pooling data from grades 1, 4 and 6. We used these three grades because they all used a common test—the Stanford Achievement Test—and were normed to a common scale. Our final Ed Tech analysis sample for the RD replication exercise consisted of 7,569 students and 355 teachers.

Table III.1 presents summary statistics for the main variables we used from the Ed Tech data, based on our analysis sample. The data used in this table, and throughout our analyses, were weighted to account for nonresponse and unequal probabilities of assignment to treatment.[28] Exactly half of all sample members were female, and their mean age was 9.55 years. The average scores for the pretests and posttests shown in the table are Stanford Achievement Test scores aggregated for grades 1, 4, and 6, measured in normal curve equivalent (NCE) units. The NCE scale is a transformation of percentile units to an equal interval scale, with a mean of 50 and standard

[25] The second year report includes impact estimates for the separate software products, but these were not estimated for the first year report.

[26] Parent consent was obtained for 93 percent of all students in these classrooms (Dynarski et al. 2007).

[27] This sample included only students who completed both a pretest and posttest. Students who moved into or out of study schools during the study year were excluded. Dynarski et al. (2007) found that the impact estimates based on a sample of all students who had spring test scores (regardless of whether or not they had fall, or baseline, scores) had the same sign and significance as the main impact estimates presented in the report.

[28] The probability of assignment to treatment was higher in school-grade blocks where an odd number of teachers participated in the study. We constructed weights to equalize the influence of treatment and control observations within each block. In particular, we constructed a base weight for each sample member using the inverse of the probability of assignment to the intervention group to which he or she was assigned. The base weight was then multiplied by non-response weights (included in the original Ed Tech data set) to produce the final weights.

deviation of 21 in a nationally representative norm group of children in the same grade. Although we pooled sample members in the three grades, students' test scores reflect their performance relative to other students in their same grade nationally. The Ed Tech sample looked reasonably similar to the national student population for these grades, with the mean pretest NCE score for the sample of 47.2 only slightly below the national mean of 50. Teachers in the sample had 10.7 years of experience teaching, on average, and 88 percent were female. Means for the treatment and control groups are also reported separately in the table. There were no significant differences between the two groups for any of the baseline variables, based on two-tailed t-tests of the equality of means.

Table III.1. Characteristics of the ED Tech Experimental Sample

	Full Sample		Control		Treatment	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Student Characteristics						
Proportion Female	0.50	0.50	0.51	0.50	0.49	0.50
Age (Years)	9.55	2.15	9.57	2.16	9.53	2.15
Proportion In Treatment Classroom	0.51	0.50	0.00	0.00	1.00	0.00
Pretest Score[a]	47.23	20.71	47.61	20.94	46.86	20.47
Posttest Score[a]	48.30	20.23	47.97	20.04	48.62	20.41
Teacher characteristics						
Proportion Female	0.88	0.33	0.91	0.29	0.85	0.36
Years of Teaching Experience	10.71	9.15	10.81	9.38	10.61	8.94
Sample Size						
Students	**7569**		**3168**		**4401**	
Teachers	**355**		**157**		**198**	

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original Ed Tech study. There were no statistically significant differences between the mean values for the treatment group versus the control group.

[a] Test scores are reported in NCE units. The test has a mean of 50 and SD of 21.06.

During the design phase of this replication exercise, we were sensitive to the possibility that some of our modeling choices in estimating impacts using the RD design could have been influenced by having advance knowledge of the experimental impact estimates. To minimize this risk, we specified—but did not actually estimate—the experimental impact estimates we aimed to replicate before beginning our analysis. For this phase, we pooled grades 1, 4, and 6 in the Ed Tech data, so we did not replicate the specific impact estimate reported in Dynarski et al. (2007). By doing this, we were able to conduct the RD analysis before estimating the target RA impacts.

B. Regression Discontinuity Results

This section presents the basic RD analysis for the Ed Tech dataset, focusing primarily on the RD High sample but with both RD High and RD Low estimates presented at the end of the section. As discussed above, students' pretest scores served as the assignment variable in the RD analysis. In the RD High sample, students whose pretest scores were above the median received the treatment, while those whose pretest scores were below the median did not receive the treatment and were in the control group.[29] We used grade-specific medians, so that there were an approximately equal number of treatment and control students within each of the three Ed Tech grade levels included in the analysis. We also "re-centered" the data so that each test score in the final dataset is reported relative to the grade-specific cutoff. For each observation, we subtracted the grade-specific median from the original NCE score. After this re-centering, zero is the cutoff for all grades.

The resulting RD High data set is summarized in Table III.2.[30] The statistics are similar to those shown previously for the whole sample, with the exception of the test scores. Because the treatment and control groups were determined entirely by a student's pretest score, the two groups had very different test score profiles. The control group, which included only observations with pretest scores below the median, had a mean pretest score of 33.5 points. In contrast, the mean pretest score for the treatment group was 66.5. The re-centered scores are also included in this table, since we use the re-centered scores in our analysis. The difference between the treatment and control groups for all test scores is significant at the 0.01 level, but the other variables remain balanced across groups.

Figure III.1 shows students' treatment status graphed against their pretest scores, using the re-centered scores. This figure shows that the dataset we created mimics a strict RD assignment rule. Students whose pretest scores were above the cutoff received the treatment and students whose scores were below the cutoff did not.

Figure III.2 presents the density graph that corresponds to the McCrary (2008) test that examines the integrity of the assignment variable. Each circle represents the density of observations falling within that cell of the assignment variable (that is, the corresponding range of values of the pretest score). The dotted and solid lines are fitted values showing the smoothed mean of the density in the cells around each value of the assignment variable (local polynomial approximations)

[29] Attrition is calculated from the pretest to the posttest. In other words, it captures cases with valid pretest data that have missing posttest data. For the Ed Tech study, the attrition rate from pretest to posttest was 4%.

[30] The corresponding table for the RD Low data set is shown in Appendix Table A.1. Subsequent tables summarizing the detailed analysis for the RD Low data set are also included in Appendix A. These results are referenced in this chapter where appropriate.

on each side of the cutoff point. The densities on each side of the cutoff are similar, with the local polynomial approximations nearly meeting at the cutoff and the associated test statistic is -0.30, which is not statistically significant. Thus, there is no evidence that the assignment variable was manipulated to affect the treatment assignment of individuals. Because we have constructed the RD dataset artificially from experimental data using a strict assignment rule, we have institutional evidence that the assignment variable was not manipulated. We explore the consequences of manipulation in Chapter 6.

Figure III.3 graphs posttest scores against pretest scores. To construct this graph, and the ones that follow, we first created "bins" that represent small intervals of values of the assignment variable, baseline test scores. We then calculated the average value of the outcome (posttest scores) for observations within each bin of the assignment variable. Each point in Figure III.3 represents an average outcome value within a given bin. The bins are equally sized at 0.5 points and begin on either side of the cutoff point, allowing us to see whether there is a visually significant jump in the outcome at the cutoff. If there is a treatment effect, we would expect to see a discontinuity in the outcome measure at the cutoff. We also conducted a formal statistical test for such a discontinuity based on the regression results that follow.

Table III.2. Characteristics of the Ed Tech RD High Sample

	Full Sample		Control		Treatment	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Student Characteristics						
Proportion Female	0.50	0.50	0.49	0.50	0.51	0.50
Age (Years)	9.50	2.09	9.48	2.10	9.52	2.10
Proportion In Treatment Classroom	0.51	0.50	0.00	0.00	1.00	0.00
Pretest Score[a]	50.45	20.38	33.51	11.36	66.50**	12.53
Re-centered Pretest Score[b]	0.45	20.38	-16.49	11.36	16.50**	12.53
Posttest Score[a]	48.73	20.58	34.51	14.32	62.20**	16.04
Re-centered Posttest Score[b]	0.70	20.34	-13.46	14.30	14.11**	15.54
Teacher characteristics						
Proportion Female	0.88	0.33	0.91	0.29	0.85	0.36
Years of Teaching Experience	10.66	9.15	10.64	9.33	10.64	9.0
Sample Size						
Students	**3681**		**1484**		**2197**	
Teachers	**348**		**194**		**194**	

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original Ed Tech study.

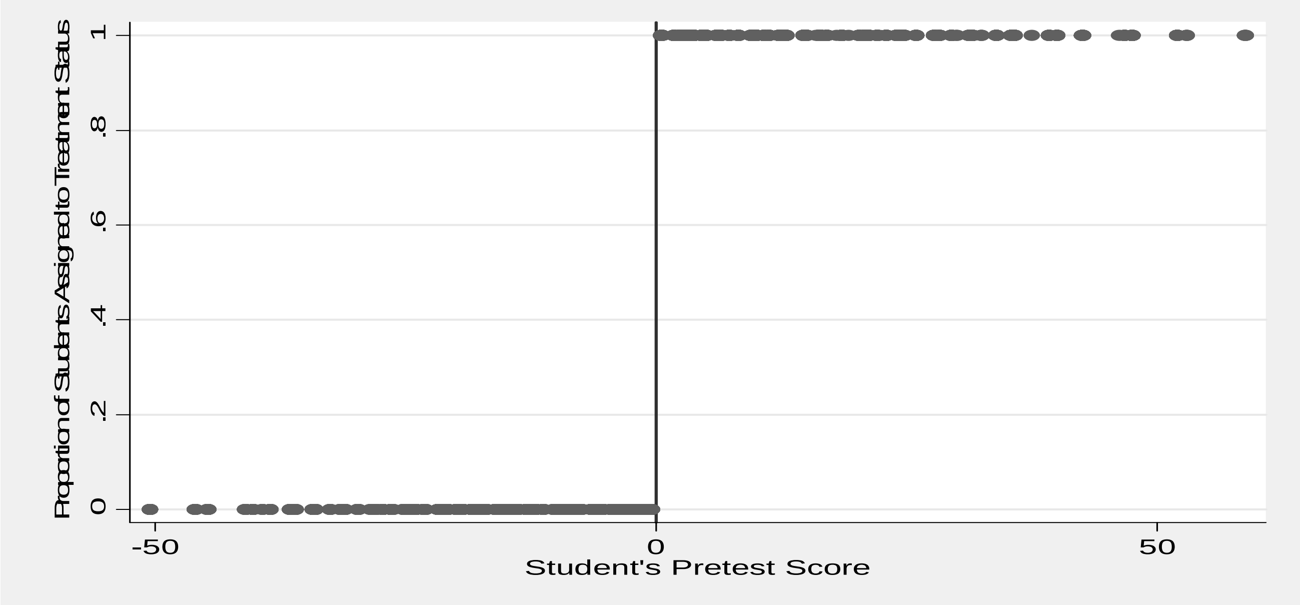
[a] Test scores are reported in NCE units. The test has a mean of 50 and SD of 21.06.

[b] Re-centered test scores are calculated by subtracting the grade-specific median from the original test score. This results in an RD cutoff score of zero for all grades.

* Difference between treatment students and control students is significantly different from zero at the .05 level, two-tailed test.

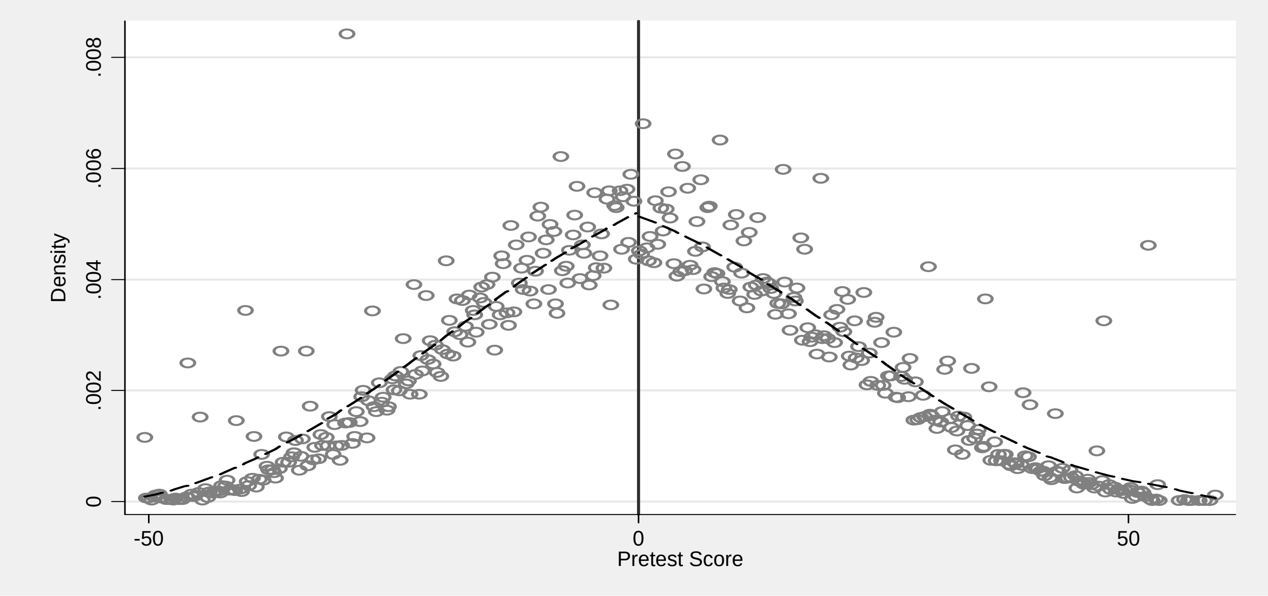
** Difference between treatment students and control students is significantly different from zero at the .01 level, two-tailed test.

Figure III.1. Probability of Assignment to the Treatment Intervention, by Students' Pretest Score: Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure III.2. Density of Pretest Scores: Ed Tech RD High Sample



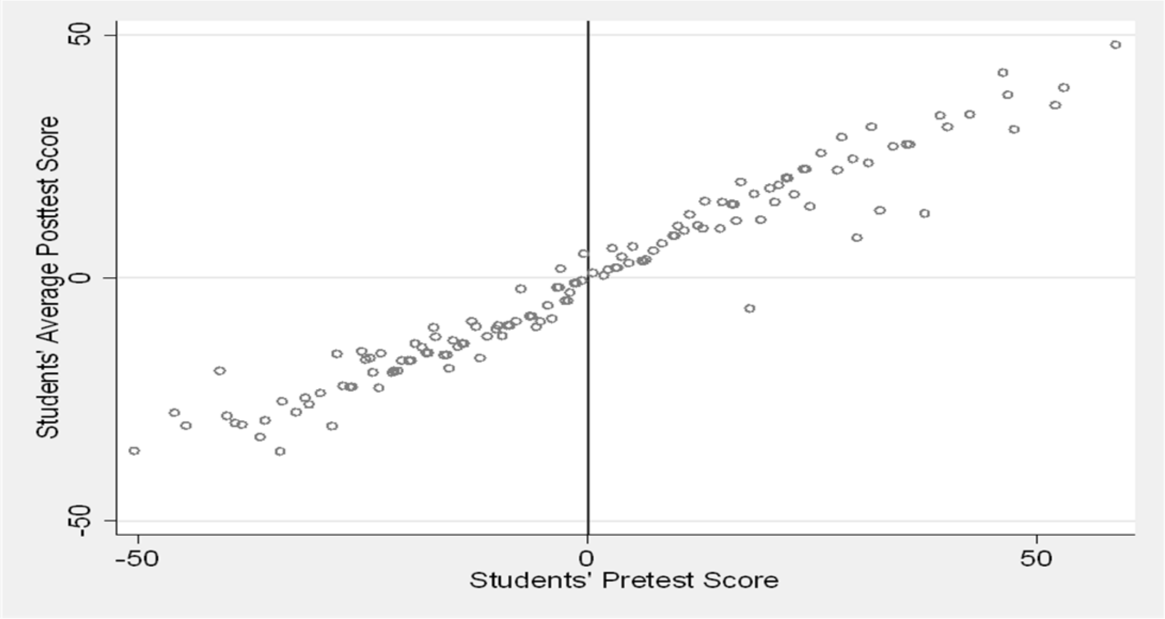
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Estimate of discontinuity at cutoff = -0.01

McCrary Test Z-stat = -0.30

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

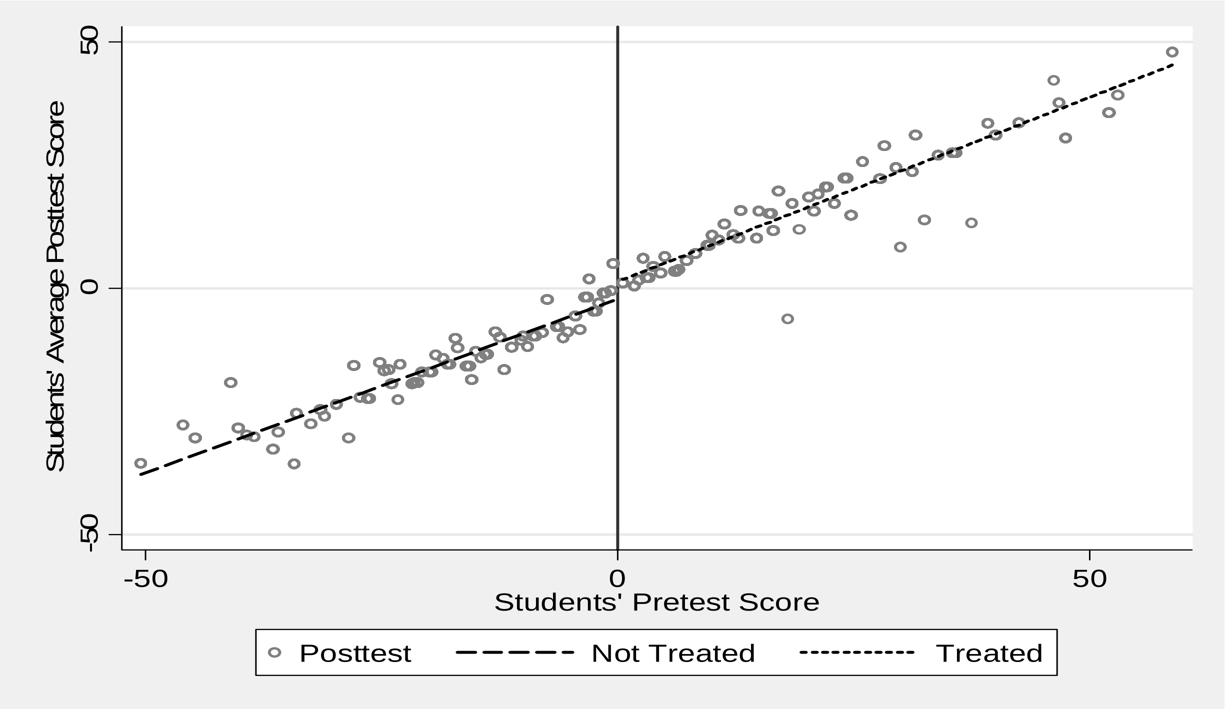
Figure III.3. Scatterplot of Students' Posttest Scores, by Pretest Score: Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

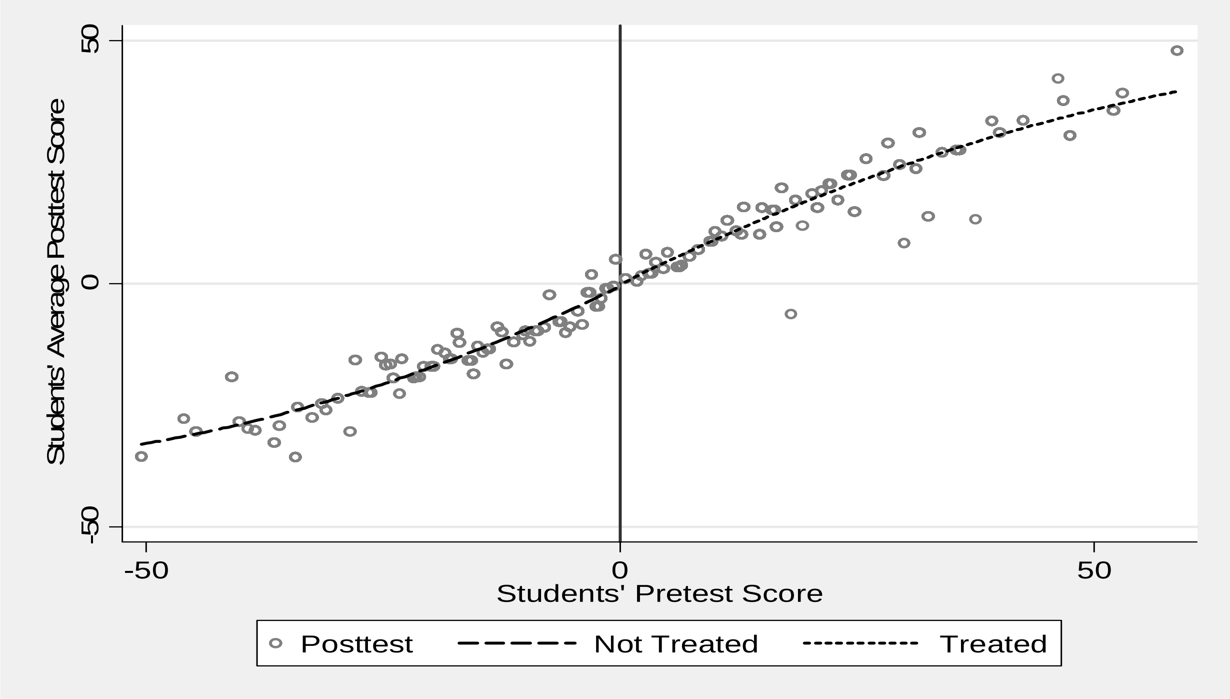
In estimating regression discontinuity models, the choice of specification is critical. As described in Chapter II, we estimated both nonparametric (local linear) and parametric RD models, using an optimal bandwidth selection procedure for the local linear models and a pre-specified procedure for choosing the appropriate parametric model. Imbens and Lemieux (2008) recommend local linear models because of their favorable bias properties at boundaries, but many education researchers choose to use parametric models in order to preserve a larger sample size. Figures III.4 to III.6 present the parametric fit to the data for linear, quadratic and cubic models. These figures highlight the subjective nature of graphical RD analysis. Our goal was to use the model that fit the data best at the cutoff point, but based on this graphical evidence it was difficult to say with any confidence which of these models best fit the data. While the relationship between the pretest (assignment variable) and posttest (outcome) appeared to be close to linear in Figures III.3 and III.4, it also seemed plausible that the linear model missed some curvature in the relationship that the quadratic and cubic models picked up. As shown in the following table, the quadratic fit was the optimal model chosen by our selection procedure, which sequentially tested the significance of each higher-order term.

Figure III.4. Relationship between Students' Pretest and Posttest Scores, Linear Specification: Ed Tech RD High Sample



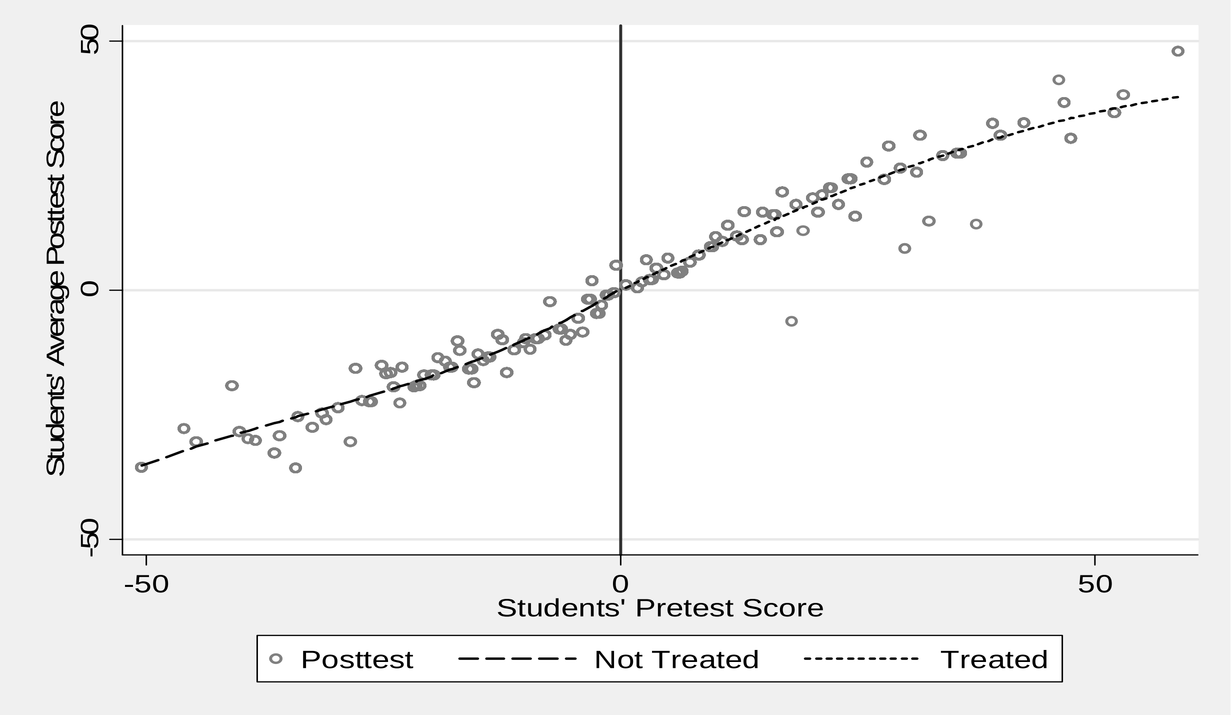
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure III.5. Relationship between Students' Pretest and Posttest Scores, Quadratic Specification: Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure III.6. Relationship between Students' Pretest and Posttest Scores, Cubic Specification: Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Table III.3 presents the estimated impacts of the Ed Tech intervention from these parametric RD models (based on results for the RD High sample). The first column shows the regression results for the linear parametric model. The second column adds squared terms that allow a quadratic relationship between the assignment variable and outcome both above and below the cutoff. The p-value for the F-test that tests the joint significance of these added terms is less than 0.005, indicating that these terms do have predictive power in the model. The third column adds cubic terms to test whether the cubic model fits the data better. The p-value for these terms is 0.61, so we chose the quadratic model as our preferred specification. The estimated treatment effect in this preferred quadratic specification is -0.10, and not statistically significant.[31] In other words, using a parametric RD model, we found that the ED Tech intervention did not significantly affect student test scores. If we had used the linear model, the incorrect parametric specification, we would have concluded that the impact was statistically significant and positive at 3.12 NCE points.

[31] The RD Low estimates presented in Table A.2 also suggest a quadratic specification, but the estimated treatment effect in that model is positive (2.82) and statistically significant.

Table III.3. Estimated Impact of Treatment Status on Test Scores, Regression Discontinuity Parametric Specifications—Ed Tech

	Specification 1	Specification 2	Specification 3
RD High Model	Linear	Quadratic	Cubic
Treatment Status	3.12** (0.78)	-0.10 (1.00)	-0.233 (1.27)
Pretest	0.70** (0.03)	1.00** (0.08)	1.12** (0.18)
Pretest * Treatment	0.02 (0.04)	-0.077 (0.10)	-0.280 (0.23)
Pretest Squared		0.007** (0.00)	-0.014 (0.01)
Pretest Squared * Treatment		-0.012** (0.00)	-0.014 (0.01)
Pretest Cubed			0.0001 (0.000)
Pretest Cubed * Treatment			-0.00 (0.00)
Sample Size	**3681**	**3681**	**3681**
R-squared	**0.69**	**0.70**	**0.70**
F-test p-value for squared terms		**0.00**	
F-test p-value for cubed terms			**0.61**

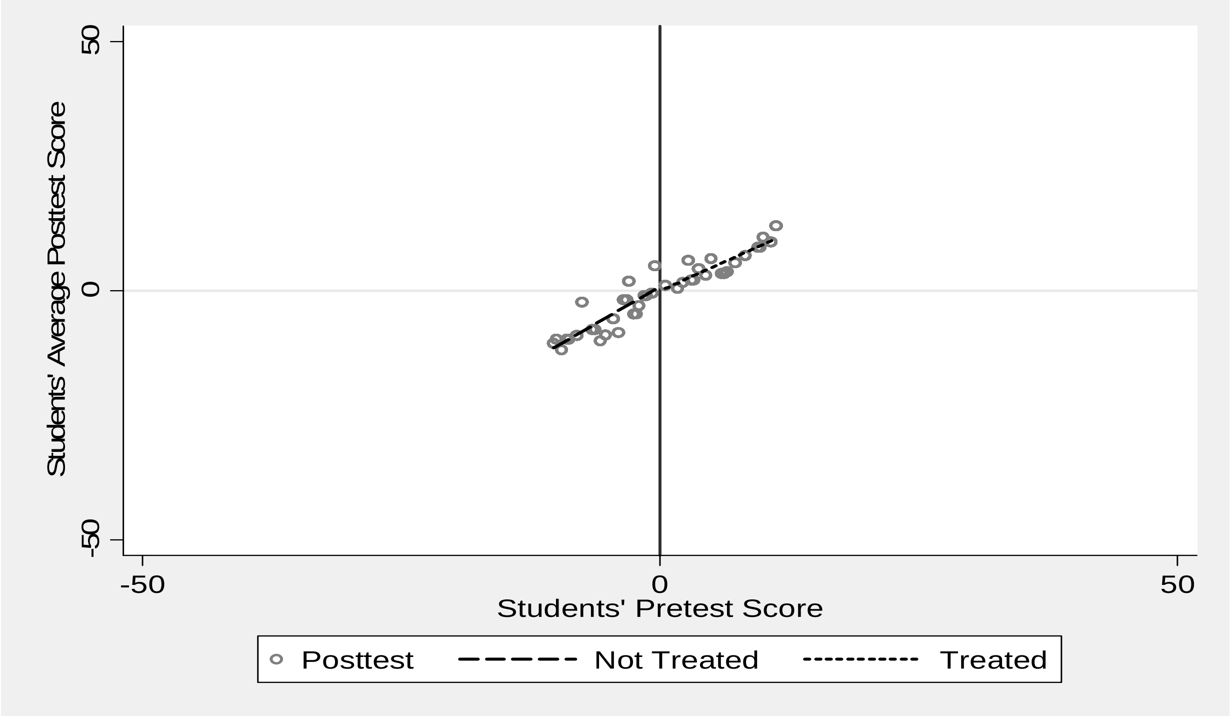
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included block fixed effects and teacher random effects. Standard errors are shown in parentheses. Specification 2 is the preferred model based on the F-tests presented at the bottom of the table.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

In the local linear (non-parametric) model, we used the algorithm described by Imbens and Kalyanaraman (2009) for selecting the optimal bandwidth.[32] For the RD High sample, the optimal bandwidth was 11.48 points.[33] Figure III.7 shows the local linear fit using only data within the optimal bandwidth and the local linear regression line fit to the data. As is customary in the RD literature, we also report results using bandwidths of one-half and twice this optimal bandwidth as a sensitivity check. Figures III.8 and III.9 show the local linear fit using these alternate bandwidths.

Figure III.7. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (Optimal Bandwidth): Ed Tech RD High Sample

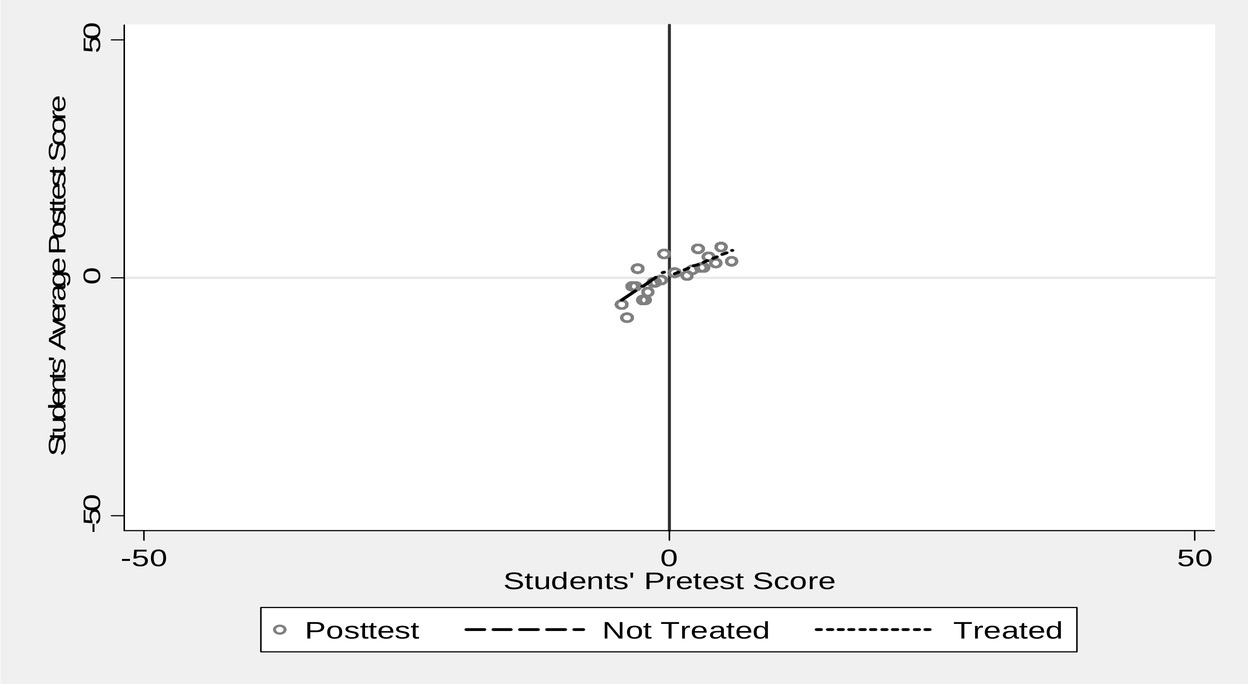


Source: Data from the Educational Technology Study (Dynarski et al. 2007).

[32] As described in Chapter 2, we also used a cross-validation procedure to choose an optimal bandwidth to test the sensitivity of our results to the bandwidth selection procedure. The cross-validation procedure did not result in a unique bandwidth choice, but the optimal bandwidth chosen with the IK procedure was within the range of bandwidths suggested by the cross-validation procedure.

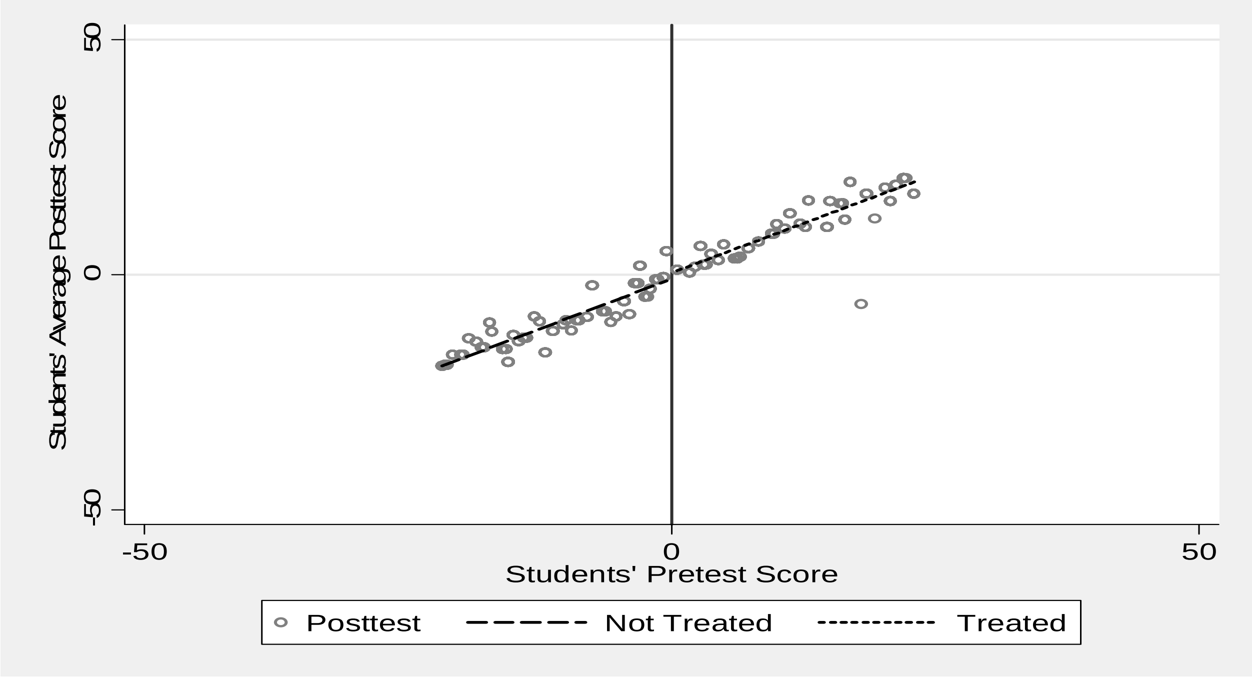
[33] The sample size using the optimal bandwidth was 1513 observations, implying that the optimal bandwidth of 11.48 points on either side of the cutoff value captured 41 percent of the full RD High sample.

Figure III.8. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (One-Half Optimal Bandwidth): Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure III.9. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (Two Times Optimal Bandwidth): Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Table III.4 presents the estimated impacts of the Ed Tech intervention using the nonparametric local linear specification with the RD High sample. Specification 1, using the optimal bandwidth (our preferred specification), yielded a point estimate of the impact of the Ed Tech intervention of -1.14, but was not statistically significant.[34] Under specifications 2 and 3, using the alternate bandwidths, the estimated impact of the Ed Tech intervention was slightly different but also not significantly different from zero.

Table III.4. Estimated Impact of Treatment Status on Test Scores, Regression Discontinuity Nonparametric Specifications—Ed Tech

	Specification 1	Specification 2	Specification 3
RD High Model	Optimal Bandwidth	(1/2)*Optimal Bandwidth	2*Optimal Bandwidth
Treatment Status	-1.14 (1.27)	-2.90 (2.06)	0.73 (0.93)
Pretest	1.16** (0.14)	1.59** (0.51)	0.86** (0.05)
Pretest * Treatment	-0.27 (0.18)	-0.79 (0.61)	-0.04 (0.07)
Sample Size	**1513**	**765**	**2684**
R-Squared	**0.36**	**0.30**	**0.54**
Bandwidth Size	**11.48**	**5.74**	**22.97**

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included block fixed effects and teacher random effects. Standard errors are shown in parentheses.

*Significantly different from zero at the .05 level, two-tailed test.

**Significantly different from zero at the .01 level, two-tailed test.

[34] In the RD Low analysis, the optimal bandwidth was found to be 14.34 and the estimated treatment effect was positive (1.57) but not statistically significant.

Table III.5 reports the results of two sets of specification checks that test the continuity of the outcome-assignment variable relationship assumption of the RD model. The first set of columns reports the results of a set of RD models in which the posttest (outcome) variable has been replaced as dependent variable with various baseline characteristics. Thus, each column shows an estimate of the discontinuity at the cutoff value in the relationship between the assignment variable and a particular baseline covariate. These baseline variables cannot have been influenced by the treatment, so we would expect their relationship with the assignment variable to be continuous at the cutoff value. If we were to find significant discontinuities in these relationships, we would be concerned that something unrelated to the treatment assignment was changing at the cutoff value and could be causing spurious impact estimates in our main RD model. Of the four reported estimates, none was statistically significant at the .05 level.

The second set of specification checks, shown in the last two columns Table III.5, shows the results of checks for spurious discontinuities in the relationship between the assignment variable and outcome at points other than the cutoff. An underlying assumption of the RD model is that, absent the treatment, the relationship between the assignment variable and outcome is continuous at the cutoff. If we found discontinuities in the relationship at points other than the cutoff, we would be concerned about the smoothness of the underlying relationship at the cutoff itself. There was no evidence of discontinuities in this relationship at the 40th or 60th percentile.

Table III.5. Regression Discontinuity Specification Checks, Nonparametric Specifications Ed Tech

	Discontinuity at the Cut-Point in Relationship Between Baseline Covariates (Other than the Assignment Variable) and the Assignment Variable				Discontinuity in Assignment Variable-Outcome Relationship at Points Other than Cut-Point	
RD High Model	% Female	% Black	% Hispanic	Age	40th Percentile	60th Percentile
Estimated Discontinuity	-0.034 (.053)	-0.013 (.038)	0.034 (.032)	-0.015 (.053)	0.23 (1.20)	1.72 (1.99)

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. Estimates shown in the table are based on local linear RD using the optimal bandwidth calculated for this dataset. The models using alternate cut-points include block fixed effects and teacher random effects. Standard errors are shown in parentheses. None of the estimates was found to be statistically significantly different from zero at the 0.05 level, two-tailed test.

Overall, using an RD model to estimate the impact of the Ed Tech intervention with the RD High sample, we found that the impact on students' test scores was negative but not statistically significant. In the parametric specification, our protocol led us to choose the quadratic model, and the impact estimate was -0.10. In the non-parametric local linear model, the optimal bandwidth was 11.48, and the estimated impact was -1.14.

Table III.6 presents a summary of the RD Low and RD High estimates. In the case of the parametric specification, the analysis of the two samples—which are independent of one another—lead to somewhat different results. The estimated impact is positive and statistically significant in the case of the RD Low sample, but close to zero and not statistically significant in the case of the RD High sample. There is only modest overlap in the 95% confidence intervals of the two estimates. In the case of the local linear specification, the impact estimates from both samples are not statistically significant, and there is more overlap in their 95% confidence intervals.

Table III.6. Summary of RD High and RD Low Impact Estimates—Parametric and Nonparametric Ed Tech Specifications

Sample/Specification	Preferred Specification/ Optimal Bandwidth	Impact Estimate	95% Confidence Interval
Parametric			
RD High Sample	Quadratic	-0.10 (1.00)	-2.06 to 1.86
RD Low Sample	Quadratic	2.82** (0.96)	0.94 to 4.70
Nonparametric (Local Linear)			
RD High Sample	11.48	-1.14 (1.27)	-3.63 to 1.35
RD Low Sample	14.34	-1.57 (1.06)	-0.51 to 3.65

Source: Data from the Educational Technology Study (Dynarski et al. 2007.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

IV. RD ESTIMATION OF THE IMPACT OF TFA

In addition to replicating experimental estimates with an RD model using the Ed Tech data, we repeated the replication exercise with data from a second study. Replicating (or failing to replicate) experimental impact estimates in a different context provides additional evidence about the performance of RD methods and greater statistical power for the comparison of the two sets of estimates. We used data from the Teach for America (TFA) Study (Decker et al. 2004) for this second replication effort. Like the Ed Tech study, the TFA study used an experimental design to estimate impacts, and provided a reasonably large sample of about 1,800 students for the comparison of RD impact estimates with the original experimental impact estimates.

This chapter describes the results of the estimation of impacts based on an RD design using data from the TFA study. The first section describes the TFA study and associated data, and the second section presents the RD estimates of the TFA impact using the RD High sample. RD estimates of the impact of TFA based on the RD Low sample are presented in Appendix A.

A. Description of Study and Data

Teach for America (TFA) was founded in 1989 to address the educational inequities facing children in low-income communities by expanding the pool of teacher candidates available to the schools those children attend. TFA recruits from the nation's top colleges among those seniors and recent graduates who are willing to commit to teach for a minimum of two years in low-income schools. TFA teachers have strong academic records, but limited training in pedagogy. The TFA study examined the impact of TFA teachers on students in their classrooms, compared with what would have happened in the absence of the TFA teachers (Decker et al. 2004).

To ensure the comparability of students in TFA classrooms and students in non-TFA classrooms, the study randomly assigned students to classrooms before the start of the school year. Students randomly selected into the treatment group entered the classroom of a TFA teacher, while those selected into the control group entered the classroom of a non-TFA teacher. The schools in the study were chosen to be broadly representative of schools where TFA placed teachers. The study included schools in 6 of the 15 TFA regions, selected after stratifying on urbanicity and the dominant race/ethnicity of students served by the schools. The study team focused on elementary schools, grades 1-5, because these elementary classes are structured to be similar within grade and students are assigned to a single homeroom teacher who teaches both reading and math. For each eligible school, all grades with at least one TFA teacher and one control teacher were included in the study.

The random assignment of students was conducted within school and grade level blocks. The final analytic sample included 1,765 students in 37 random assignment blocks across 17 schools. There were 44 TFA teachers and 785 students in TFA classrooms. The control group included 56 teachers and 980 students in non-TFA classrooms.

The TFA study team administered the Iowa Test of Basic Skills (ITBS) as a pretest in the fall and a follow-up posttest in the spring.[35] Students in both treatment and control classrooms completed the assessments.[36] Table IV.1 presents the demographic characteristics and pretest scores of the treatment and control groups. The data used in this table, and throughout our analyses, were weighted to account for nonresponse and unequal probabilities of assignment to treatment.[37] The analysis of baseline data suggested that random assignment was successful, with no significant demographic or baseline achievement differences between students in the treatment and control classrooms (Table IV.1). The study team continued to monitor classroom rosters throughout the year to guard against violations of the random assignment design.

On average, students in the TFA study sample were disadvantaged and low achieving. Over 95 percent of the students in the sample were eligible for free or reduced-price school lunch benefits. The pretest scores of the students in the study are far below the level of children in the same grade nationally. The mean score for students in the TFA sample was 27 Normal Curve Equivalent (NCE) points in mathematics and 26 in reading, substantially below the national mean of 50.

As in the ED Tech analysis, we were sensitive to the possibility that some of our modeling choices in estimating impacts using the RD design could have been influenced by having advance knowledge of the experimental impact estimates. To minimize this risk, we specified—but did not actually estimate—the experimental impact estimates we aimed to replicate before beginning our analysis. The original impact estimates in Decker et al. (2004) estimated a separate treatment effect for each random assignment block and aggregated these estimates conditional on grade, school, and district level controls. In this replication exercise, we estimated a more parsimonious specification, which allowed us to conduct the RD analysis before estimating the target experimental impacts. The following section reports the results of our RD analysis.

[35] Attrition from the pretest to posttest for the treatment and control students was 11.1% for Math and 9.7% for Reading.

[36] Nine percent of the students were in classrooms that provided instruction in Spanish. In these classrooms, the assessments were administered in Spanish.

[37] The sample weights correct for two features of the study design. The first of these features was sample attrition, as eleven percent of sample members did not complete the spring assessment. The non-response weights were higher for students with characteristics similar to the attrited students. The second feature was that the probability of assignment to treatment varied by block. This variation occurred because while the majority of blocks had one TFA teacher and one control teacher, others might have had one TFA teacher and two control teachers or two TFA teachers and one control teacher. To correct for this arbitrary treatment-block correlation, each sample member was given a weight proportional to the inverse of the sample size within the block of the treatment group to which he/she was assigned.

B. Regression Discontinuity Results

We structured the RD analysis of TFA impacts in a similar way as the analysis of Ed Tech impacts presented in Chapter III. We used students' pretest scores as the assignment variable, so that in the RD High sample students whose pretest scores were above the median received the treatment, while those whose pretest scores were below the median were in the control group. We used grade-specific medians to ensure that the grade distribution in the RD sample mirrors the grade distribution in the original experimental sample. We implemented the RD design separately for our analyses of math and reading test scores.[38]

Table IV.1. Characteristics of the TFA Experimental Sample

	Full Sample		Control		Treatment	
Student Characteristics	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Proportion Female	0.49	0.50	0.49	0.50	0.50	0.50
Proportion Black	0.65	0.48	0.63	0.48	0.67	0.47
Proportion Hispanic	0.29	0.46	0.29	0.46	0.29	0.45
Age (Years)	8.5	1.6	8.5	1.6	8.5	1.6
Proportion Free Lunch	0.81	0.39	0.79	0.40	0.82	0.38
Pretest Math Score[a]	31.8	17.9	32.3	18.1	31.3	17.7
Pretest Reading Score[a]	33.5	19.4	32.7	19.5	34.2	19.2
Posttest Math Score[a]	30.7	18.6	31.1	19.2	30.2	18.0
Posttest Reading Score[a]	31.6	20.1	31.4	19.9	31.8	20.3
Sample Size						
Students	**1642**		**925**		**717**	
Teachers	**100**		**56**		**44**	

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original TFA study. There were no statistically significant differences between the mean values for the treatment group versus the control group.

[a] Test scores are reported in NCE units. The test has a mean of 50 and SD of 21.

The resulting RD High data set for math is summarized in Table IV.2. As with the Ed Tech data, we re-centered all test scores so that they are relative to the grade-level median. All of the treatment students had pretest scores that exceeded their grade-level median and all of the control students had pretest scores below the median. Thus, there is a large and statistically significant difference between mean pretest scores for treatment versus control students (44.9 for the treatment group and 21.1 for the control group). In the RD High sample, there were also statistically

[38] One implication of this strategy was that an individual student in the TFA study sample could have been in the treatment group for the RD High analysis for math but not in the treatment group in the RD High analysis for reading.

significant demographic differences between treatment and control students. Treatment students were less likely to be black (61 percent of the treatment group and 68 percent of the control group) and more likely to be Hispanic (33 percent of the treatment group and 27 percent of the control group).

Table IV.2. Characteristics of the TFA Math RD High Sample

	Full Sample		Control		Treatment	
Student Characteristics	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Proportion Female	0.50	0.50	0.47	0.50	0.52	0.50
Proportion Black	0.65	0.48	0.68	0.47	0.61*	0.49
Proportion Hispanic	0.30	0.46	0.27	0.44	0.33*	0.47
Age (Years)	8.6	1.6	8.6	1.6	8.7	1.5
Proportion Free Lunch	0.81	0.40	0.81	0.39	0.80	0.40
Pretest Math Score[a]	31.9	17.6	21.1	12.8	44.9**	13.2
Pretest Reading Score[a]	33.5	19.5	25.4	17.6	43.2**	17.1
Posttest Math Score[a]	31.5	18.2	25.2	17.7	39.1**	15.9
Posttest Reading Score[a]	31.8	19.6	25.4	18.0	39.4**	18.8
Sample Size						
Students	**807**		**494**		**313**	
Teachers	**99**		**55**		**44**	

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original TFA study. One control classroom did not complete the math pretest and is excluded from the RD math samples.

[a] Test scores are reported in NCE units. The test has a mean of 50 and SD of 21.

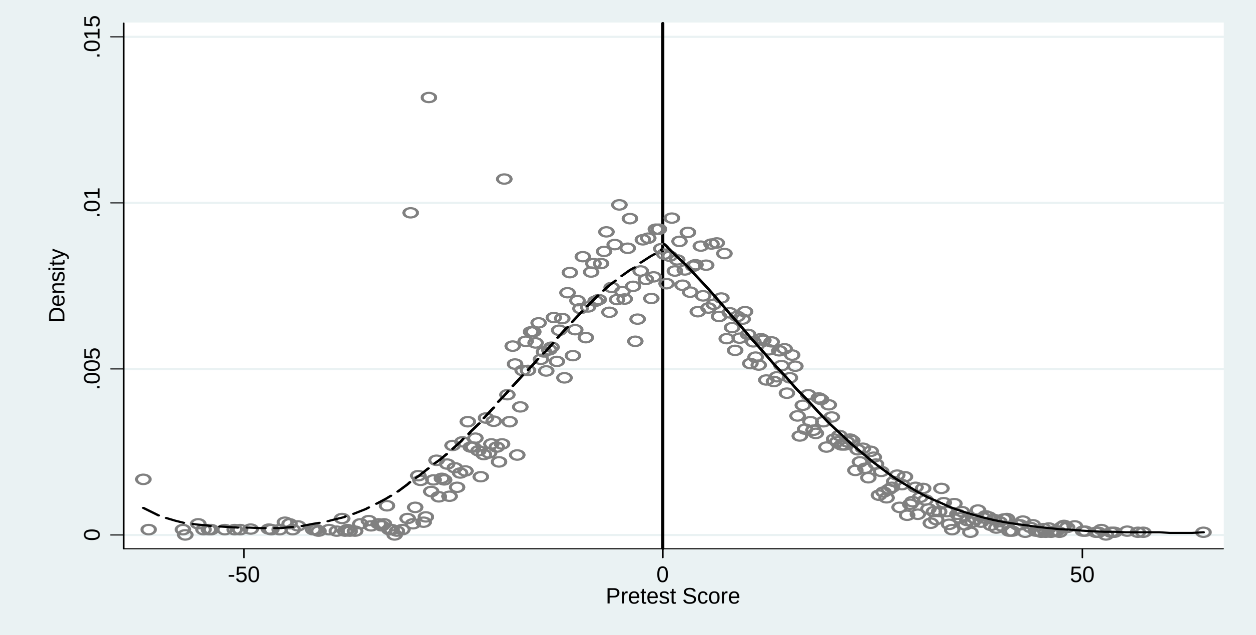
* Difference between treatment students and control students is significantly different from zero at the .05 level, two-tailed test.

** Difference between treatment students and control students is significantly different from zero at the .01 level, two-tailed test.

1. Math

Any RD analysis depends on the integrity of the assignment variable—here, students' math pretest score. Figure IV.1 presents the density graph that corresponds to the McCrary test, the statistical test of the assignment variable's integrity. Each circle represents the density of observations falling within that cell of the assignment variable. The densities on each side of the cutoff are similar, with the local polynomial approximations nearly meeting at the cutoff and the associated test statistic is 0.26, which is not statistically significant. Thus, there is no evidence of inappropriate assignment. Since we constructed the RD dataset artificially from experimental data using a strict assignment rule, we also have institutional evidence that the assignment variable was not manipulated.

Figure IV.1. Density of Pretest Scores in TFA Math RD High Sample



Estimate of discontinuity at cutoff = 0.02

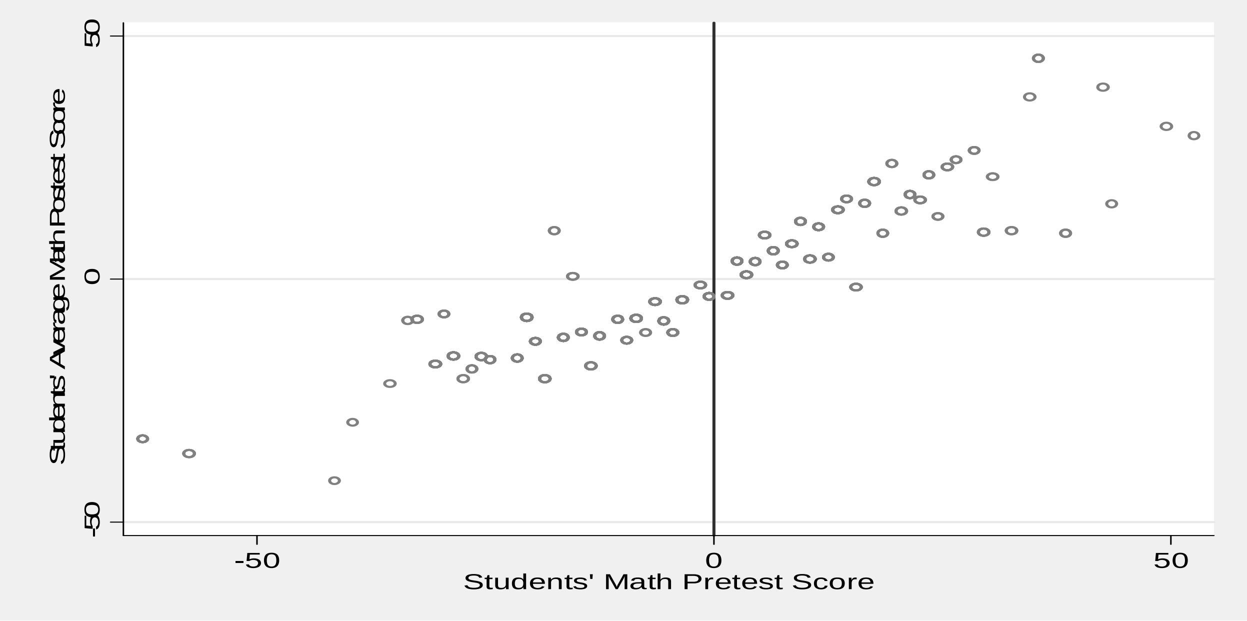
McCrary Test Z-stat = 0.26

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

Figure IV.2 shows the relationship between the assignment variable in the RD analysis—students' math pretest score—and the outcome variable—students' math posttest score. To construct this graph, and the ones that follow, we again used "bins" that represent small intervals of the baseline test score, and calculated the mean value of the outcome (posttest scores) for observations within each bin of the assignment variable. Each point in Figure IV.1 represents an average outcome value within a given bin. The bins are equally sized at 1 point.[39] The relationship between the pretest and posttest score appears linear near the cutoff, but there is more variation in the tails of the test score distribution. From the scatter plot, it is difficult to discern if there is a discontinuity in the outcome measure at the cutoff.

Figure IV.2. Scatterplot of Students' Posttest Scores, by Pretest Score: TFA Math RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

[39] These bins are larger than the 0.5 point bins used in the Ed Tech analysis because the TFA sample size is substantially smaller. Using the smaller 0.5 point bins would have resulted in more cases in which mean values for a given bin were based on few observations.

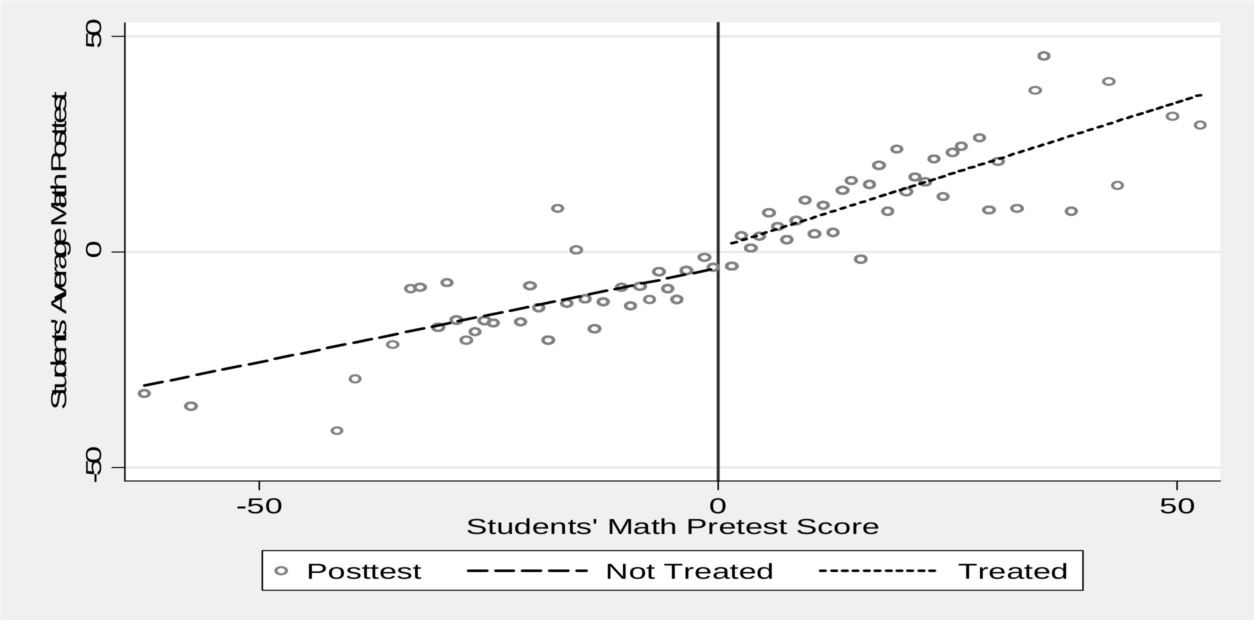
Figures IV.3 to IV.5 present a series of parametric specifications of the assignment variable-outcome relationship, including linear, quadratic, and cubic fits. In each, we interacted the linear, quadratic, and cubic terms with the treatment indicator to allow for different relationships above and below the cutoff. To select one of these specifications for our primary parametric model, we followed the pre-determined rule described in Chapter II. We added higher order pretest terms until the joint test of the higher order term and its interaction with the treatment variable was not statistically significant.

For the Math RD High sample, the quadratic specification was the preferred estimate – the quadratic terms were significant and the cubic terms were not. In this specification, the estimated impact of the TFA program on math test scores was 1.31 points, but the estimate was not statistically significant (Table IV.3). The table also shows the estimated impacts from the linear and the cubic specifications. These results highlight the sensitivity of the estimated impacts to the specification choice. If the relationship between the math pretest and posttest had been modeled as linear, the estimated TFA impact would have been positive and statistically significant (4.78 points). If the relationship had been modeled as cubic, the estimated impact would be negative and not significant.

As an alternative to the parametric RD design, we also estimated non-parametric local linear specifications that used a restricted set of data near the cutoff. Based on the IK procedure, the optimal bandwidth we selected was 14.1 points. By restricting the analysis to students within this bandwidth, the local linear analysis used 64 percent of the student observations.

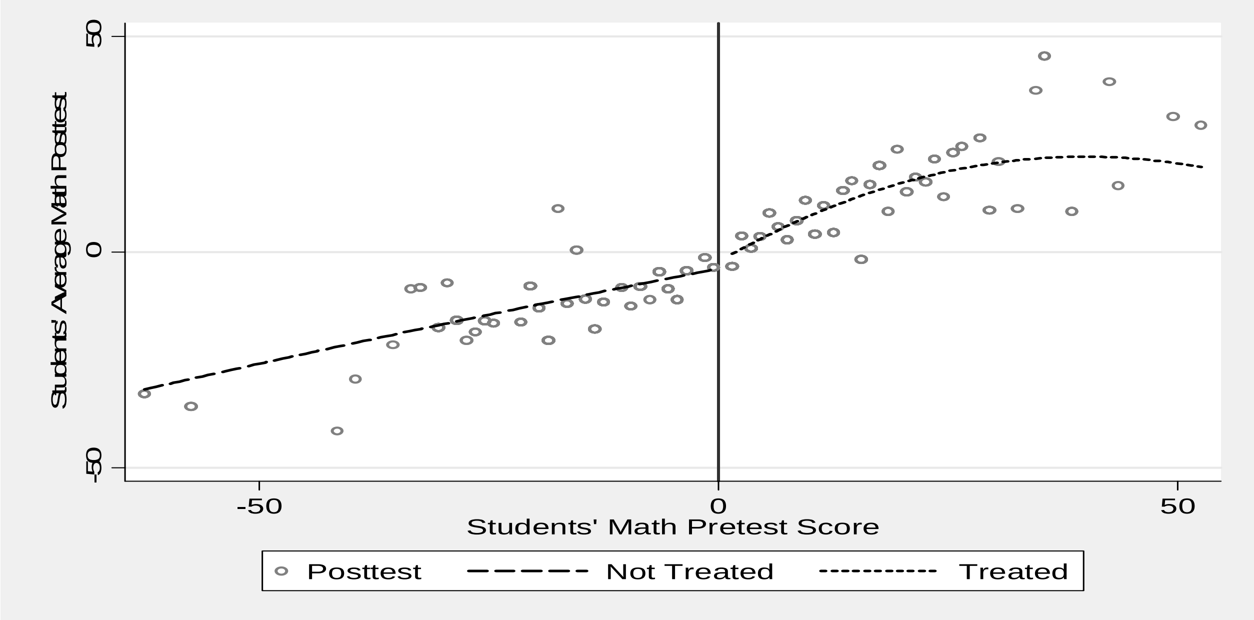
Within the optimal bandwidth, the relationship between the pretest and posttest was assumed to be linear. In Figure IV.6, we show the local linear fit within the optimal bandwidth. With this optimal bandwidth local linear specification, the estimated impact of TFA on math scores was 0.91 points, and the estimate was not statistically significant (Table IV.4). As a sensitivity analysis, Table IV.4 also reports the estimated impact using half of the optimal bandwidth and twice the optimal bandwidth. Figures IV.7 and IV.8 show the local linear fit using these alternate bandwidths. In both cases, the estimates were positive but not significant. The estimates do highlight the relationship between the size of the bandwidth and the variance of the impact estimate. The standard error of the estimated treatment effect decreases with the size of the bandwidth.

Figure IV.3. Relationship between Students' Pretest and Posttest Scores, Linear Specification: TFA Math RD High Sample



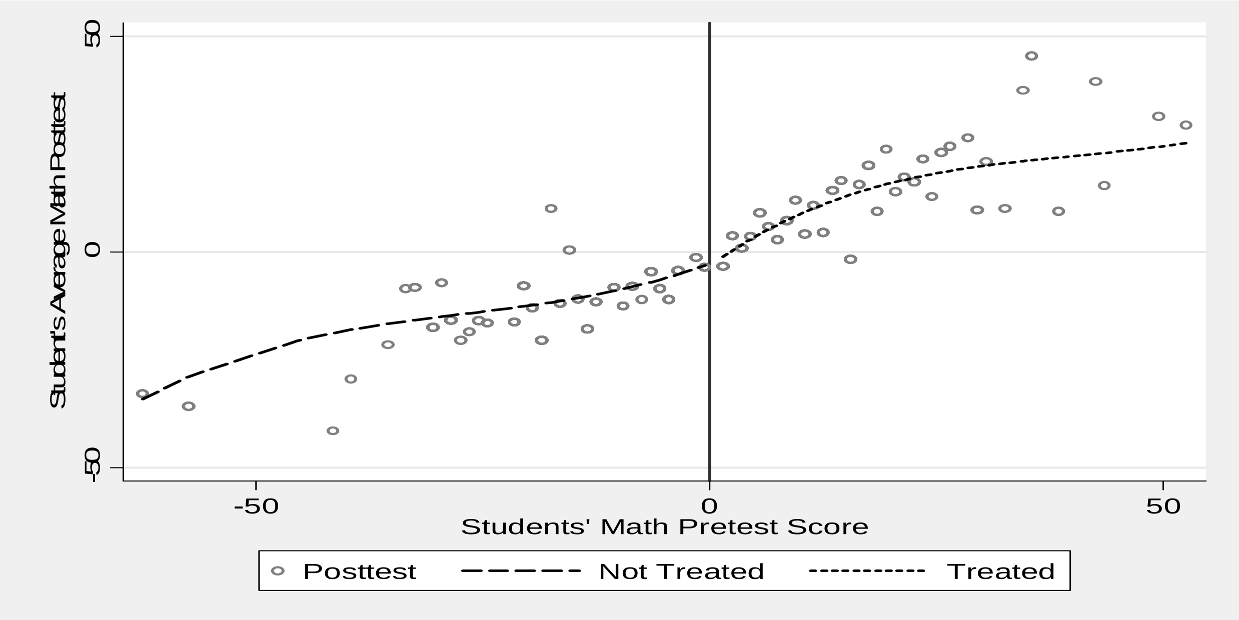
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.4. Relationship between Students' Pretest and Posttest Scores, Quadratic Specification: TFA Math RD High Sample



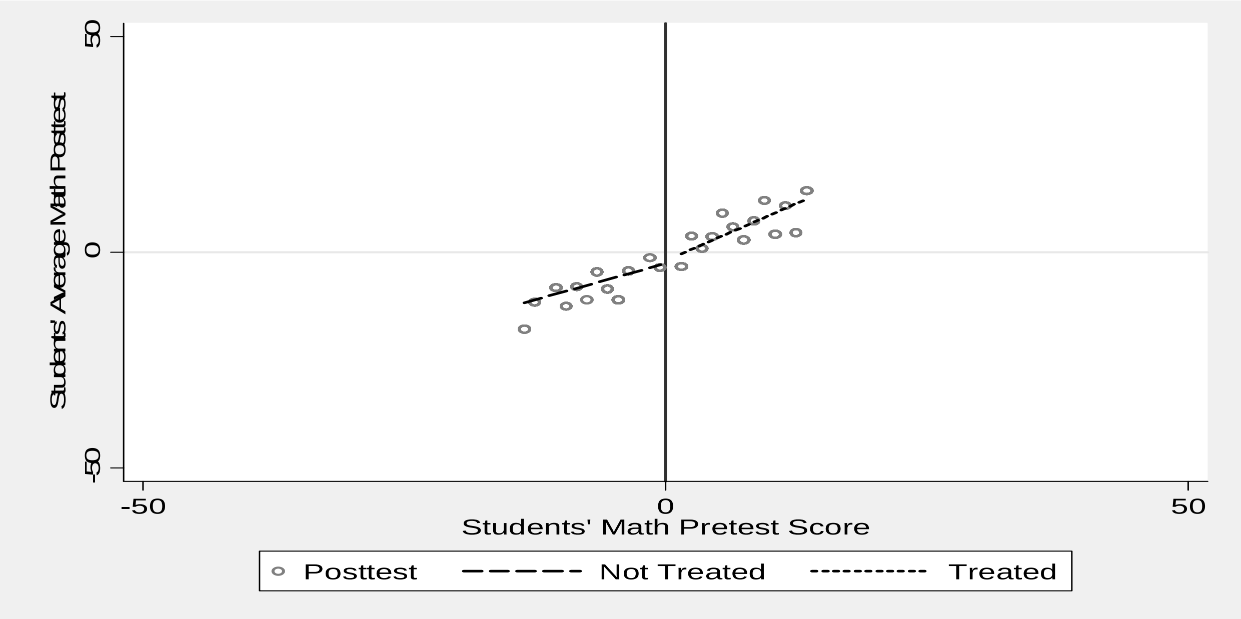
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.5. Relationship between Students' Pretest and Posttest Scores, Cubic Specification: TFA Math RD High Sample



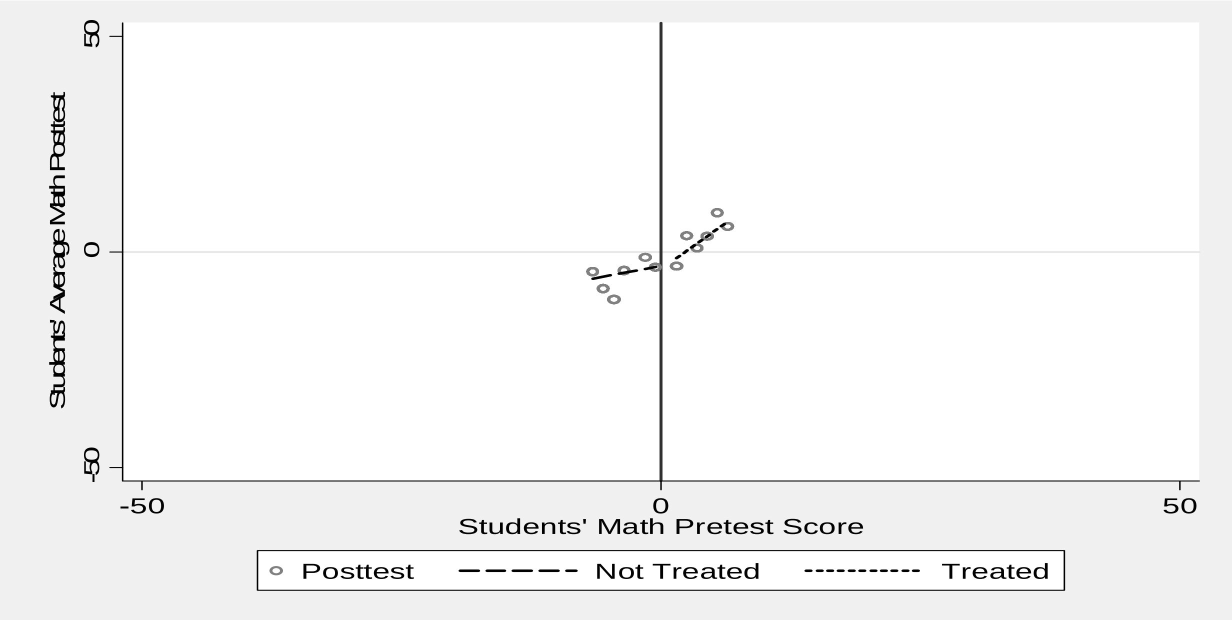
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.6. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (Optimal Bandwidth): TFA Math RD High Sample



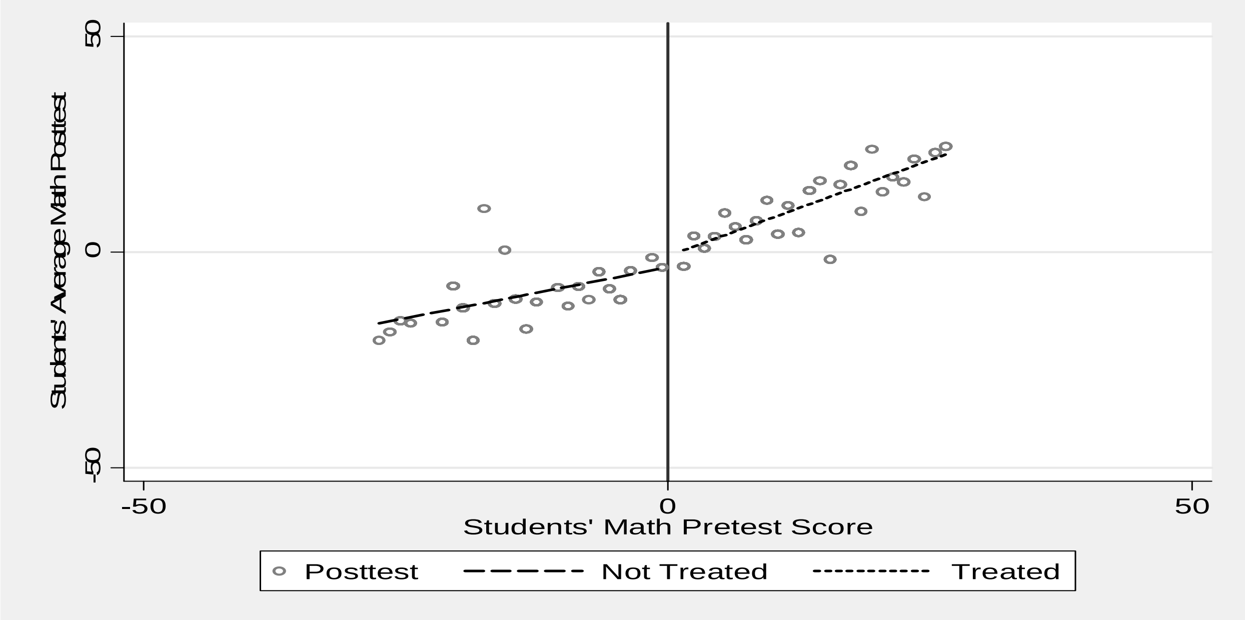
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.7. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (One-Half Optimal Bandwidth): TFA Math RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.8. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (Two Times Optimal Bandwidth): TFA Math RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Table IV.3. Estimated Impact of Treatment Status on Math Test Scores, Regression Discontinuity Parametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD High Model	Linear	Quadratic	Cubic
Treatment Status	4.78** (1.57)	1.31 (2.04)	-0.41 (2.71)
Pretest	0.41** (0.06)	0.42** (0.13)	0.68* (0.31)
Pretest * Treatment	0.27** (0.10)	0.86** (0.26)	0.86 (0.58)
Pretest Squared		0.00 (0.00)	0.01 (0.02)
Pretest Squared * Treatment		-0.02** (0.01)	-0.05 (0.03)
Pretest Cubed			0.00 (0.00)
Pretest Cubed * Treatment			0.00 (0.00)
Sample Size	**804**	**804**	**804**
R-squared	**0.46**	**0.46**	**0.46**
F-test p-value for squared terms		**0.02**	
F-test p-value for cubed terms			**0.54**

Source: Data from the Teach For America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect. Standard errors are shown in parentheses. Specification 2 is the preferred model based on the F-tests presented at the bottom of the table.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

In both the optimal parametric and non-parametric specifications, we found positive but not statistically significant impacts of TFA on math test scores.[40] These RD estimates relied on the assumption that the underlying relationship between the assignment variable (the student's pretest score) and outcome (the posttest score) would be continuous in the region of the cutoff value if not for the TFA treatment. We conducted the same specification tests to indirectly assess implications of this assumption that we used with the Ed Tech analysis presented in Chapter 3. First, we tested for discontinuities at the cutoff value in the relationship between pretest scores and student demographic variables including gender, race/ethnicity, and age. Restricting our analysis to students within the optimal bandwidth, there were no significant discontinuities at the cutoff value in these relationships (Table IV.5).

Table IV.4. Estimated Impact of Treatment Status on Math Test Scores, Regression Discontinuity Nonparametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD High Model	Optimal Bandwidth	(1/2)*Optimal Bandwidth	2*Optimal Bandwidth
Treatment Status	0.91 (2.18)	2.15 (3.51)	2.39 (1.74)
Pretest	0.50** (0.19)	0.12 (0.41)	0.43** (0.09)
Pretest * Treatment	0.64* (0.29)	1.16 (0.84)	0.53** (0.14)
Sample Size	**518**	**298**	**731**
R-Squared	**0.32**	**0.21**	**0.43**
Bandwidth Size	**14.1**	**7.0**	**28.2**

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect. Standard errors are shown in parentheses.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

[40] In the RD Low analysis for TFA math, the optimal parametric specification was quadratic and we found a positive and statistically significant treatment effect (4.33). In the non-parametric analysis, the optimal bandwidth was 9.0 and the estimated treatment effect was positive but not statistically significant (1.36).

We also looked for discontinuities in the relationship between the assignment and outcome variables at points other than the cutoff point. We tested for discontinuities in the math posttest at the 40th and 60th percentiles of the math pretest distribution. There was no evidence of discontinuities at either point (Table IV.5).

Table IV.5. Regression Discontinuity Specification Checks: TFA (Math Test Scores)

	Discontinuity at the Cut-Point in Relationship Between Baseline Covariates Other than the Assignment Variable and the Outcome				Discontinuity in Assignment Variable-Outcome Relationship at Points Other than Cut-Point	
RD High Model	% Female	% Black	% Hispanic	Age	40th Percentile	60th Percentile
Estimated Discontinuity	0.07 (0.09)	-0.02 (0.04)	0.06 (0.04)	-0.06 (0.11)	-1.78 (2.01)	0.70 (2.31)

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. Estimates shown in the table are based on local linear RD using the optimal bandwidth calculated for this dataset. The models using alternate cut-points include an indicator for whether the pretest was taken in Spanish, block fixed effects, and teacher random effects. Standard errors are shown in parentheses.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

2. Reading

We used the same steps to estimate the impact of TFA on reading achievement using an RD design as we used to estimate impacts on math achievement. We used the RD High Reading sample, where the treatment group included students of TFA teachers with reading test scores above their grade-specific median and the control group included students of non-TFA teachers with reading test scores below the median.

Figure IV.9 presents the density graph that corresponds to the McCrary test that examines the integrity of the reading pretest as an assignment variable. The associated test statistic is -0.39, which is not statistically significant.

Figure IV.10 shows the relationship between the reading pretest score and the reading posttest score. All reading test scores are reported relative to the grade-specific medians. As with the math achievement graphs, each point on the graph represents the average posttest score of students within a 1 point pretest score bin. The reading pretest-posttest relationship appears to be linear near the cutoff value.

Figures IV.11 to IV.13 present a series of parametric specifications of the assignment variable-outcome (pretest-posttest) relationship, including linear, quadratic, and cubic fits. We used the same decision rule for selecting an optimal parametric specification as we used for the math analysis. For reading, the optimal parametric specification was the linear fit. Unlike our parametric specification for the math achievement models, the quadratic terms were not statistically significant (Table IV.6). In the linear parametric specification, the estimated impact of TFA on reading achievement was 0.66 points, and not statistically significant. We also report the results from two other non-optimal parametric specifications in the second and third columns of Table IV.7. The estimated impact was not very sensitive to the parametric specification. The estimates from the quadratic and cubic specifications were both less than 1 point and neither was statistically significant. We also estimated non-parametric local linear specifications. The IK optimal bandwidth for the Reading RD High sample was 14.4, a bandwidth that included 59 percent of students. This bandwidth was similar in size to the optimal bandwidth for the Math RD High.[41] In Figure IV.14, we show the local linear fit within the optimal bandwidth. Using students within the optimal bandwidth, the estimated impact of TFA on reading achievement was 0.93, and the estimate was not significant (Table IV.7).[42] Figures IV.15 and IV.16 show the local linear fit using these alternate bandwidths. The impact estimate varied somewhat with alternative bandwidths (-2.92 for the specification based on half of the optimal bandwidth and -0.31 for the specification based on twice the optimal bandwidth), but none of the estimates were statistically significant.

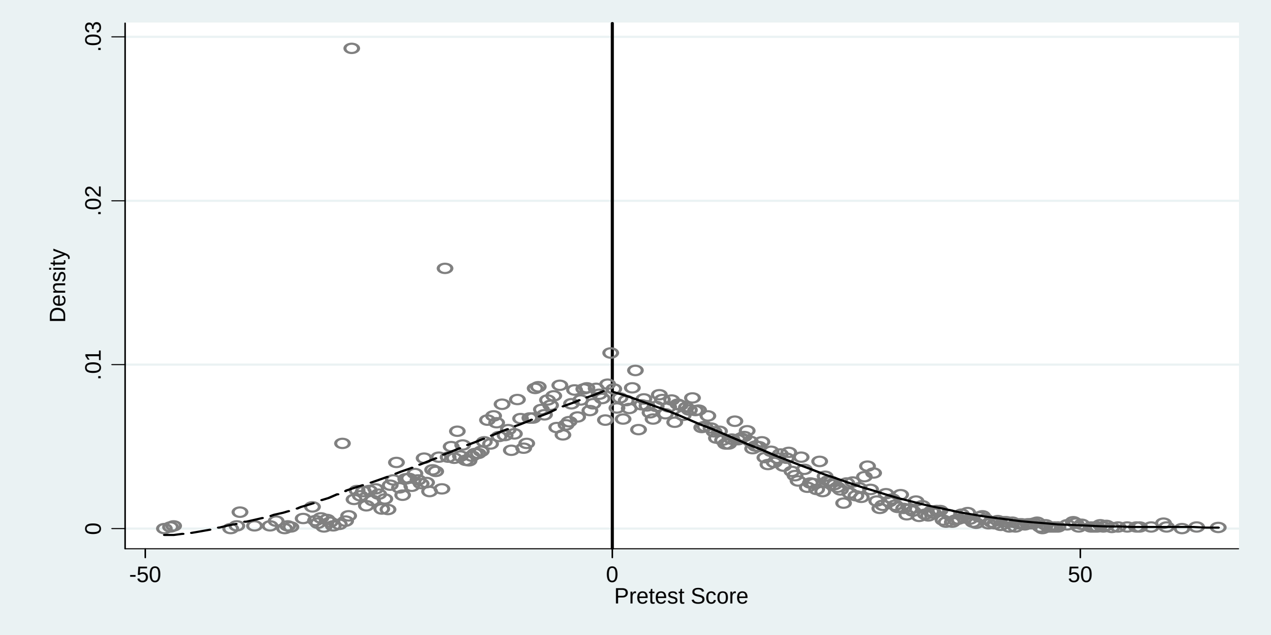
We also repeated the specification checks conducted in the analysis of math achievement to ensure that there were not discontinuities in the relationships between the reading pretest score and baseline characteristics of sample members. We did find a statistically significant discontinuity in the relationship between the reading pretest score and age, but the estimated discontinuities in the relationships between the reading pretest score and the other three characteristics were not significant (Table IV.8).[43] In the other specification check, which tested for spurious discontinuities in the relationship between the reading pretest and posttest scores at points in the distribution other than the cutoff value, we did not find a significant discontinuity at the 40th or 60th percentiles (Table IV.8).

[41] It is interesting that the optimal bandwidth was not larger for reading than it was for math given that the optimal parametric specification for reading was linear. The IK bandwidth choice may be more sensitive to small areas of non-linearity than our parametric specification choice.

[42] In the RD Low analysis for TFA reading, the optimal parametric specification was linear and we found a positive but not statistically significant treatment effect (0.75). In the non-parametric analysis, the optimal bandwidth was 12.1 and the estimated treatment effect was negative but not statistically significant (-0.60).

[43] The statistically significant discontinuity in sample members' age at the cutoff value of the assignment variable was the only case in which the specification test of the RD design "failed" across the three analyses (TFA reading, TFA math, and Ed Tech) and 18 tests. Any explanation for this significant relationship would be speculative, so we offer none. If age is a key background characteristic on which equivalence must be established for the WWC protocol under which this study would be reviewed, then this could cause the study to not meet standards.

Figure IV.9. Density of Pretest Scores in Original Dataset (TFA Read)



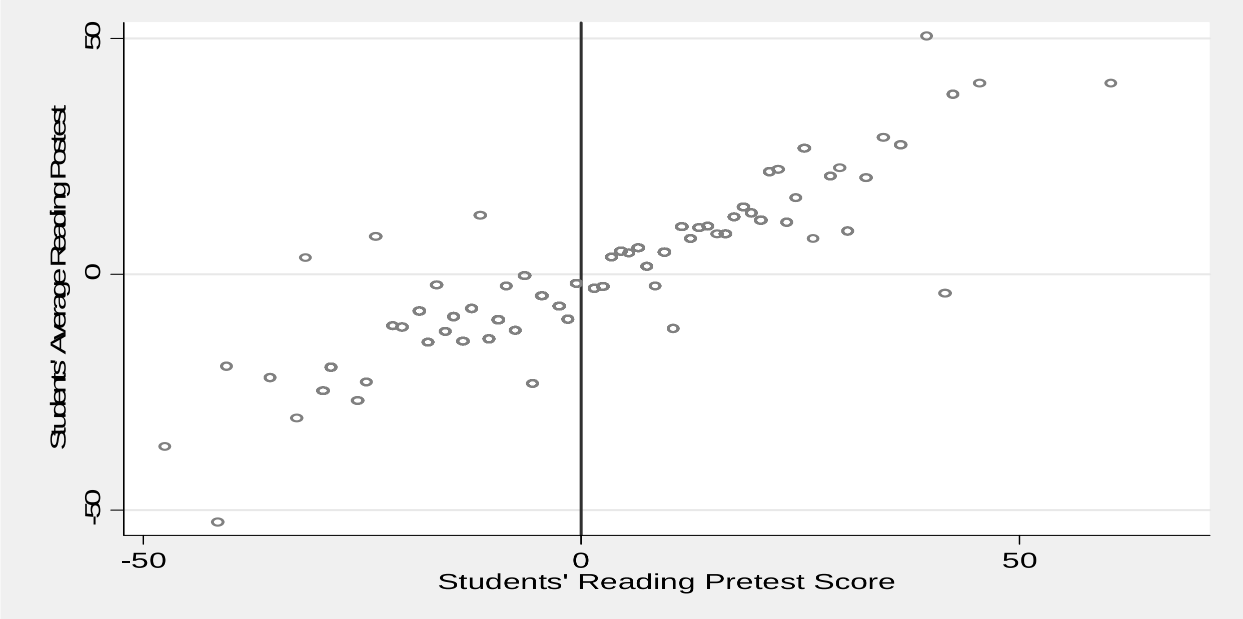
Estimate of discontinuity at cutoff = -0.03

McCrary Test Z-stat = -0.39

Source: Data from the Teach for America Study (Decker et al. 2004).

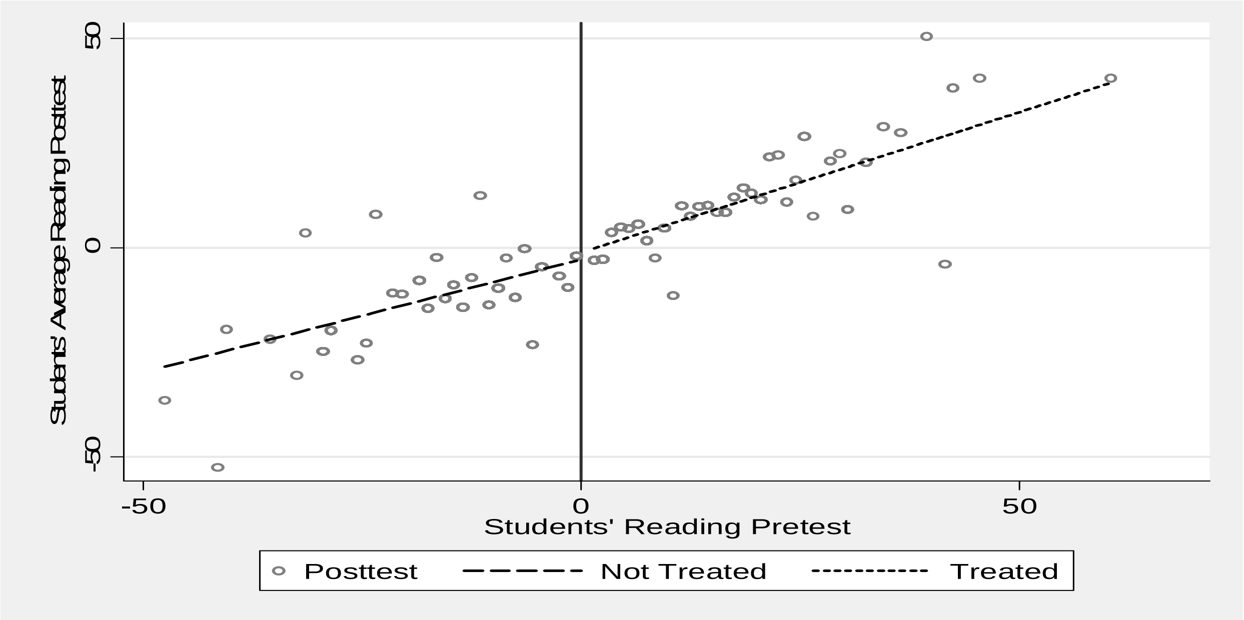
Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

Figure IV.10. Scatterplot of Students' Posttest Scores, by Pretest Score: TFA Reading RD High Sample



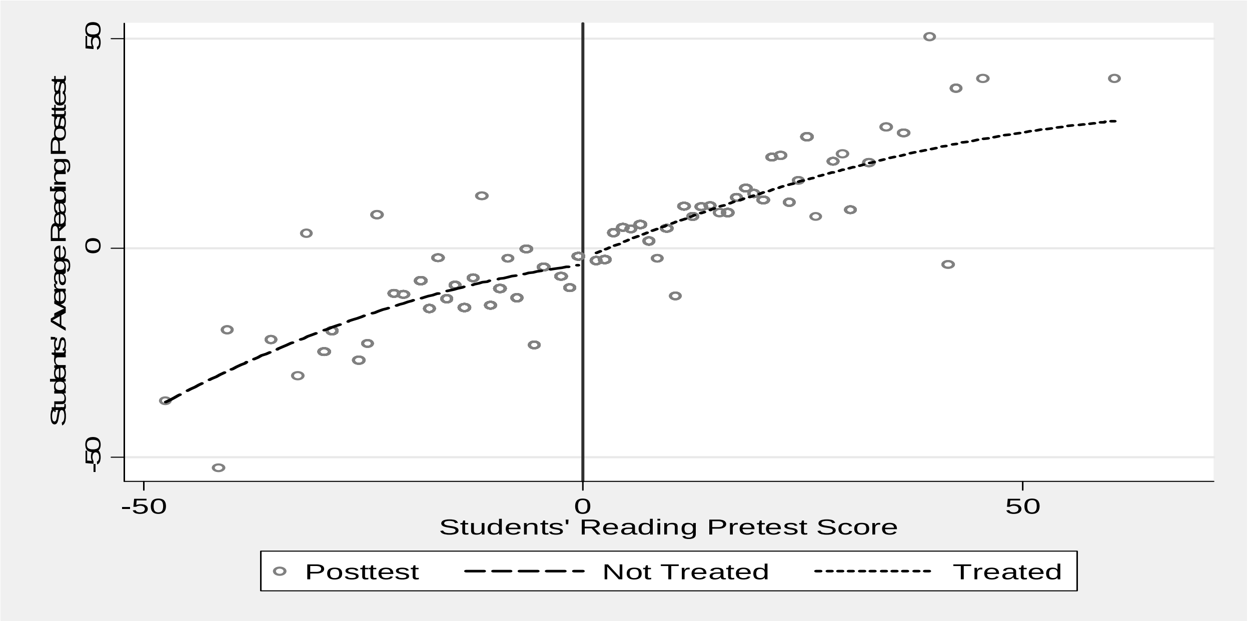
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.11. Relationship between Students' Pretest and Posttest Scores, Linear Specification: TFA Reading RD High Sample



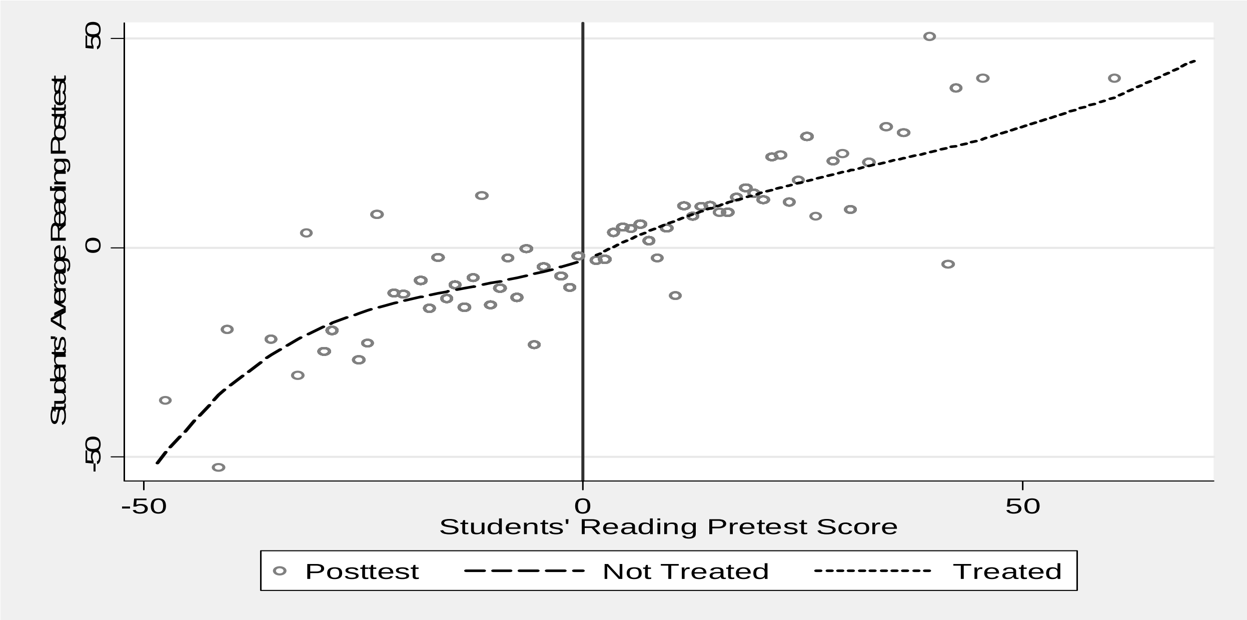
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.12. Relationship between Students' Pretest and Posttest Scores, Quadratic Specification: TFA Reading RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.13. Relationship between Students' Pretest And Posttest Scores, Cubic Specification: TFA Reading RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Table IV.6. Estimated Impact of Treatment Status on Reading Test Scores, Regression Discontinuity Parametric Specifications—TFA

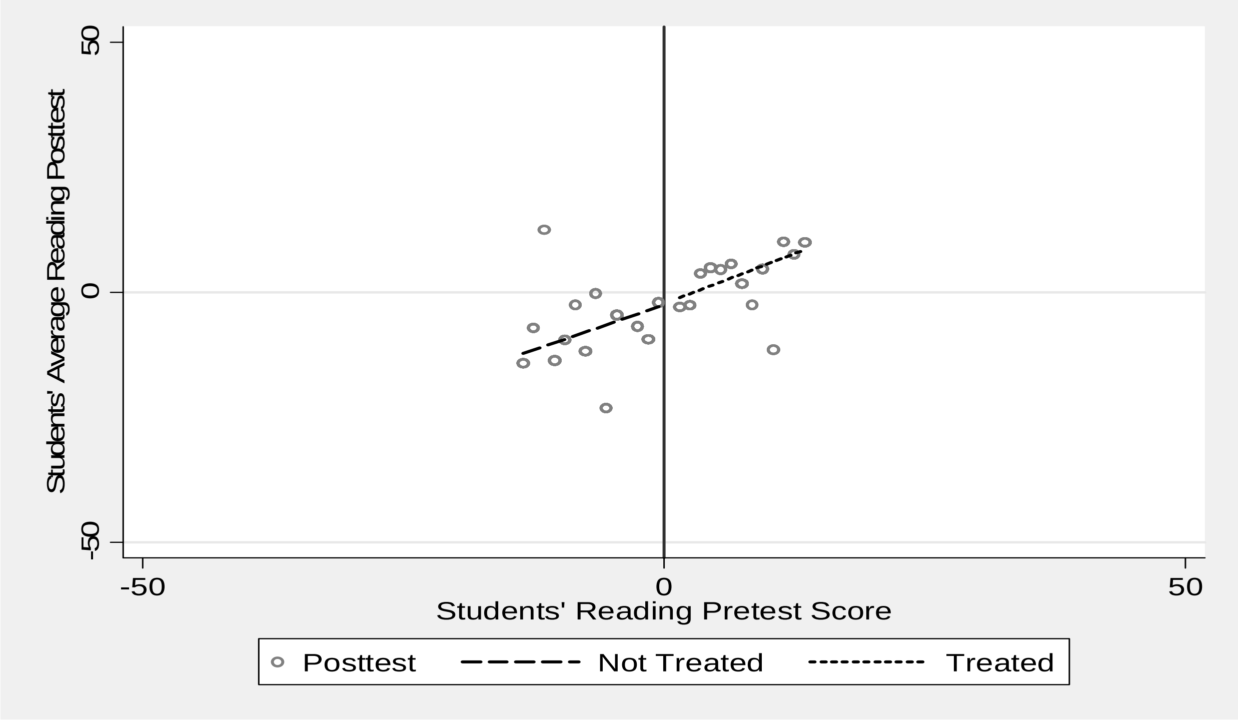
	Specification 1	Specification 2	Specification 3
RD High Model	Linear	Quadratic	Cubic
Treatment Status	0.66 (1.68)	0.14 (2.05)	0.82 (2.50)
Pretest	0.47** (0.06)	0.40* (0.18)	0.10 (0.36)
Pretest * Treatment	0.25** (0.09)	0.45 (0.26)	0.75 (0.51)
Pretest Squared		0.00 (0.01)	-0.02 (0.02)
Pretest Squared * Treatment		0.00 (0.01)	0.02 (0.03)
Pretest Cubed			0.00 (0.00)
Pretest Cubed * Treatment			0.00 (0.00)
Sample Size	**862**	**862**	**862**
R-squared	**0.49**	**0.49**	**0.49**
F-test p-value for squared terms		**0.62**	
F-test p-value for cubed terms			**0.63**

Source: Data from the Teach For America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect. Standard errors are shown in parentheses. Specification 1 is the preferred model based on the F-tests presented at the bottom of the table.

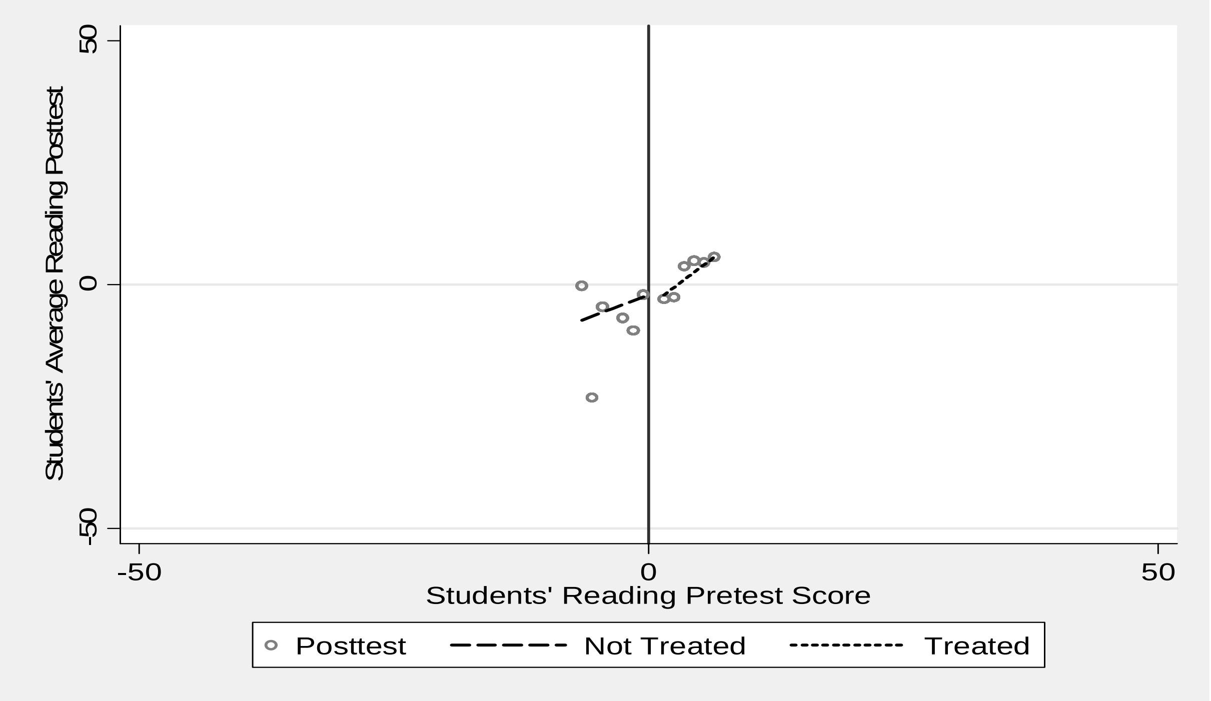
*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Figure IV.14. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (Optimal Bandwidth): TFA Reading RD High Sample



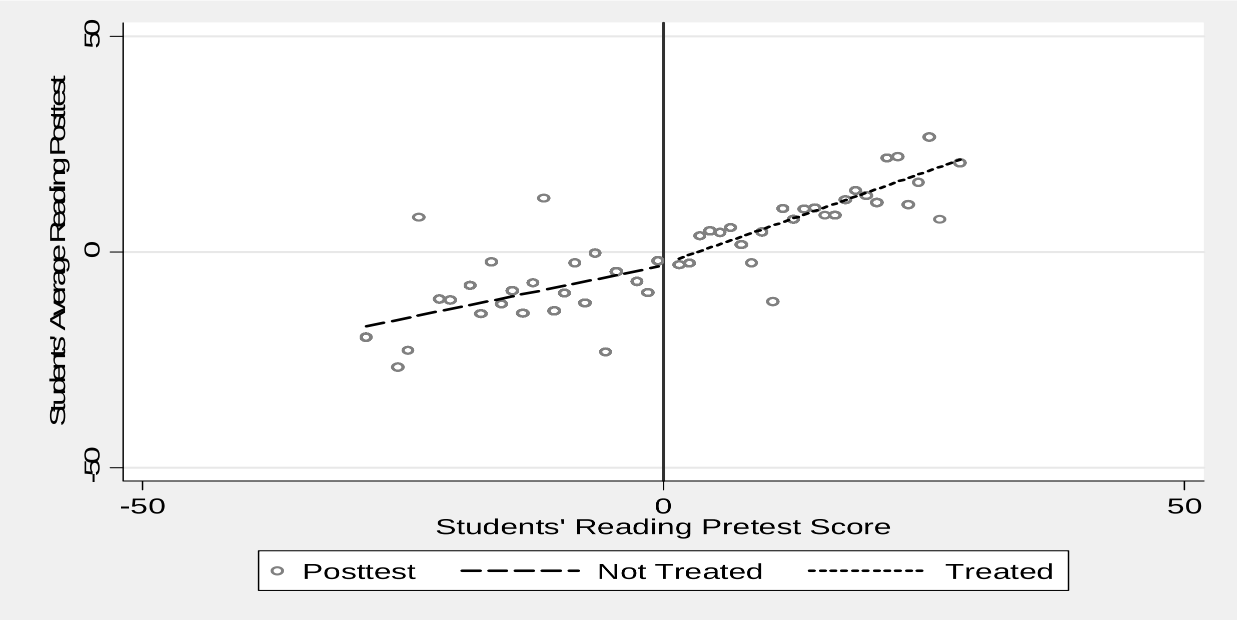
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.15. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (One-Half Optimal Bandwidth): TFA Reading RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Figure IV.16. Relationship between Students' Pretest and Posttest Scores, Local Linear Specification (Two Times Optimal Bandwidth): TFA Reading RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Table IV.7. Estimated Impact of Treatment Status on Reading Test Scores, Regression Discontinuity Nonparametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD High Model	Optimal Bandwidth	(1/2)*Optimal Bandwidth	2*Optimal Bandwidth
Treatment Status	0.93 (2.20)	-2.92 (3.39)	-0.31 (1.66)
Pretest	0.53** (0.19)	0.51 (0.47)	0.50** (0.07)
Pretest * Treatment	0.12 (0.29)	0.86 (0.80)	0.33** (0.12)
Sample Size	511	287	818
R-Squared	0.38	0.40	0.46
Bandwidth Size	14.4	7.2	28.8

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect. Standard errors are shown in parentheses.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Table IV.8. Regression Discontinuity Specification Checks: TFA (Reading Test Scores)

	Discontinuity at the Cut-Point in Relationship Between Baseline Covariates Other than the Assignment Variable and the Outcome				Discontinuity in Assignment Variable-Outcome Relationship at Points Other than Cut-Point	
RD High Model	% Female	% Black	% Hispanic	Age	40th Percentile	60th Percentile
Estimated Discontinuity	0.03 (0.09)	0.08 (0.05)	-0.01 (0.04)	-0.32** (0.11)	-0.68 (2.07)	2.99 (2.39)

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. Estimates shown in the table are based on local linear RD using the optimal bandwidth calculated for this dataset. The models using alternate cut-points include an indicator for whether the pretest was taken in Spanish, block fixed effects, and teacher random effects. Standard errors are shown in parentheses.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

3. Summary of TFA RD Results

Table IV.9 presents a summary of the RD Low and RD High estimates for the TFA estimates. In the case of the parametric specification for TFA math, the estimated impact based on the RD Low sample (4.33) is statistically significant while the estimated impact based on the RD High sample (1.31) is not significant. However, the standard error estimates are large relative to the difference in the magnitudes of the two estimates and there is substantial overlap in the two 95% confidence intervals. For the local linear specification for TFA math and for both TFA reading specifications, the estimated impacts based on the RD High and RD Low samples are close to one another.

Table IV.9. Summary of RD High and RD Low Impact Estimates—Parametric and Nonparametric TFA Specifications

Sample/Specification	Preferred Specification/ Optimal Bandwidth	Impact Estimate	95% Confidence Interval
TFA Math: Parametric			
RD High Sample	Quadratic	1.31 (2.04)	-2.69 to 5.31
RD Low Sample	Quadratic	4.33* (2.00)	0.41 to 8.25
TFA Math: Nonparametric (Local Linear)			
RD High Sample	14.10	0.91 (2.18)	-3.36 to 5.18
RD High Sample	9.00	1.36 (2.59)	-3.72 to 6.44
TFA Reading: Parametric			
RD High Sample	Linear	0.66 (1.68)	-2.63 to 3.95
RD Low Sample	Linear	0.75 (1.68)	-2.54 to 4.04
TFA Reading: Nonparametric (Local Linear)			
RD High Sample	14.40	0.93 (2.20)	-3.38 to 5.24
RD Low Sample	12.10	-0.60 (2.67)	-5.83 to 4.63

Source: Data from the Teach for America Study (Decker et al. 2004).

*Significantly different from zero at the .05 level, two-tailed test.

V. COMPARING RD AND EXPERIMENTAL IMPACT ESTIMATES

The previous two chapters presented estimates of the impacts of the Ed Tech and TFA interventions based on RD designs. To assess the performance of the RD estimator in these two contexts, this chapter presents impact estimates for these interventions based on an experimental design and compares the RD and experimental estimates. The first section presents the experimental estimates from the two studies that will serve as benchmarks against which the RD estimates will be compared. The comparison of the RD and experimental estimates is presented in the next two sections, including the basic comparison and a synthesis of results across the individual comparisons presented in the chapter.

A. Experimental Estimates

Our purpose in conducting the experimental analysis of Ed Tech and TFA data was to produce estimates of the impact of these interventions that could be used as the standard against which the RD impact estimates could be compared. In order to serve as a true "gold standard," the specific impact parameter being estimated by the two methods must be the same. As discussed in Chapter II, the experimental design's average treatment effect (ATE) evaluated over the full sample is not necessarily the same as the local average treatment effect (LATE) evaluated at the median pretest score that is produced by the RD design. If the impact of the Ed Tech and/or TFA intervention is not related to students' pretest scores, the ATE and LATE will be identical and the experimental estimate of the ATE based on the full experimental sample will be the appropriate target parameter for the RD impact estimate. As described in Chapter II, we use the local experimental estimate as our benchmark estimate, though we also show how this local experimental estimate differs from the ATE based on the full sample.

1. Experimental Results: Ed Tech

Table V.1 shows the experimental impact estimates for the Ed Tech intervention. We first present a local experimental estimate for Ed Tech by using only data from students whose pretest scores were relatively close to the median test score—the cutoff value in the RD design. In particular, we included any student whose pretest score was within 12.9 points of the median. This range was chosen to correspond with the average optimal bandwidth size from the RD High and RD Low models. The resulting sample of students in this "trimmed sample" included 45 percent of the overall Ed Teach student sample. In the trimmed sample, the estimated impact of Ed Tech was 1.41 points, which was statistically significant. This translates into an effect size of 0.07 standard deviation units. [44] The ATE based on the full sample, shown in the second column of the table, was 0.72 points (an effect size of 0.04), and the difference between the local experimental estimate and the ATE was not statistically significant at the 0.05 level of statistical significance (p-value=0.10).[45,] As described above, we defined the estimated impact of the Ed Tech intervention based on the

[44] Effect sizes represent the magnitude of the estimated impact expressed in terms of the variation in the outcome measure within the study sample. In particular, the effect sizes presented here were calculated by dividing the impact estimate by the standard deviation of the outcome variable among the full sample.

[45] The experimental estimates reported in Dynarski et al. (2007) were 0.73 for grade 1, 0.41 for grade 4, and 1.43 for grade 6. None of these impacts was statistically significant.

restricted sample (1.41 points) as the target impact parameter, or benchmark, against which we would compare the RD impact estimate.[46]

Table V.1. Experimental Impact Estimates, Alternative Specifications: Ed Tech

	Local Experimental Estimate[a]	Full Sample Experimental Estimate	Difference (col A- col B)	p-value
Coefficient Estimate	1.41** (0.53)	0.72 (0.41)	0.70 (0.42)	0.10
Effect Size Estimate	0.07** (0.03)	0.04 (0.02)	0.03 (0.02)	0.10
Sample Size	3418	7569		
R-Squared	0.33	0.67		

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports the coefficient on treatment status from regression models that also included the pretest score, random assignment block fixed effects, and a teacher random effect. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[a] Restricted bandwidth is the average of the optimal bandwidth from the RD High and RD Low samples.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

2. Experimental Results: TFA

The estimated impact of TFA on math achievement based on the local experimental sample, which included 54 percent of the original sample, was also positive and statistically significant at 2.41 NCE points (Table V.2). The estimated impact based on the experimental design for the full sample was positive and statistically significant, at 2.57 NCE points, which corresponds to an effect size of 0.13, close to the LATE estimate for the trimmed sample.[47] These two estimates were relatively close in size and not significantly different from one another (p-value=0.83).

The estimated impact of TFA on reading achievement based on the experimental design for the full sample was positive but not statistically significant, at 0.52 NCE points, or 0.03 effect size units (Table V.3).[48] For the trimmed sample, covering 55 percent of the full sample, the estimated local

[46] We performed an additional investigation into possible treatment effect heterogeneity by calculating local impact estimates for different ranges of pretest scores. Using a series of regressions of equal bandwidth, but centered at different points along the distribution of pretest scores, we find that there is some variation in impact estimates using different subsamples of data. We find, however, that the vast majority of estimates for each data set fall within the 95% confidence interval for the ATE parameter presented in this chapter and that the local estimate at our cutoff value is very close to the ATE point estimate. This analysis is described in Appendix B.

[47] Our experimental estimate for math is also similar to the impact estimate of 2.4 reported in Decker et al. (2004).

[48] The reading impact estimate in Decker et al. (2004) was 0.56.

treatment impact on reading scores was 0.39, which was not significantly different from the impact estimate for the full sample (p-value=0.83).

Table V.2. Experimental Impact Estimates on Math Scores, Alternative Specifications: TFA

	Local Experimental Estimate[a]	Full Sample Experimental Estimate	Difference (col A-col B)	p-value
Coefficient Estimates	2.41** (0.83)	2.57** (0.85)	-0.16 (0.75)	0.83
Effect Size Estimate	0.12** (0.04)	0.13** (0.04)	-0.01 (0.04)	0.83
Sample Size	893	1642		
R-Squared	0.18	0.38		

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports selected coefficients from regression models that also included random assignment block fixed effects and a classroom random effect. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[a] Restricted bandwidth is the average of the optimal bandwidth from the RD High and RD Low samples.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Table V.3. Experimental Impact Estimates on Reading Scores, Alternative Specifications: TFA

	Local Experimental Estimate[a]	Full Sample Experimental Estimate	Difference (col A–col B)	p-value
Coefficient Estimates	0.39 (0.82)	0.52 (0.63)	-0.13 (0.60)	0.83
Effect Size Estimate	0.02 (0.04)	0.03 (0.03)	-0.01 (0.03)	0.83
Sample Size	**925**	**1659**		
R-Squared	**0.28**	**0.46**		

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports selected coefficients from regression models that also included random assignment block fixed effects and a classroom random effect. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[a] Restricted bandwidth is the average of the optimal bandwidth from the RD High and RD Low samples.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

B. Comparing the Experimental and RD Estimates

The final step in the replication exercise involved comparing the estimated impacts of the Ed Tech and TFA interventions presented in Chapters 3 and 4 with the experimental estimates of these same impacts presented above. Since we defined the local experimental estimates to be the "gold standard" estimates of these impacts, we assessed whether the RD estimates matched these experimental estimates in some sense. One criterion we used for determining whether the RD impact estimates matched the experimental estimates was a test of whether the difference between the two estimates was statistically significant. Given that the RD and experimental estimates were based on different estimation methodologies and used somewhat different (though overlapping) samples, one would not expect the estimates to be identical. However, if each approach generates a consistent estimate of the same impact parameter, any difference between the two should be explained by sampling variability alone. To present a fuller picture of the estimated differences between these two estimates, we also present the 95 percent confidence interval of this difference.

A limitation of relying on a significance test of the difference between RD and experimental impact estimates is that the lower the statistical power of the test, the less likely we would be to find a significant difference between the two estimates, even if such a difference existed. In other words, we risked concluding that RD and experimental designs yielded estimates that were not significantly different from one another simply because our test lacked statistical power.

Because of this limitation of our main criterion for assessing the performance of the RD estimator, we present several other criteria for comparing the RD impact estimate with the experimental estimate. These alternative criteria are relevant for policy makers interested in understanding how policy implications may have differed between the RD and experimental impact estimates we generated. For each comparison of the RD and experimental estimates, we also asked the following questions:

- Did the RD and experimental impact estimates have the same sign and general magnitude?
- Did the estimates have the same sign and level of statistical significance?
- Did the distribution of impact estimates based on the RD design resemble the distribution of impact estimates based on the experimental design?

Table V.4 shows the RD and experimental impact estimates of the Ed Tech and TFA interventions in effect size units, based on both the parametric and nonparametric (local linear) RD estimates. Estimates are shown for both the RD High and RD Low samples. For each intervention and outcome, the table shows the RD impact estimate, the local experimental impact estimate, and the difference between the two, along with the estimated standard errors and the 95% confidence interval. As described in Chapter II, we calculated the standard error of the difference between the RD and experimental impact estimates using a bootstrap process.

Table V.4. Regression Discontinuity (RD) Versus Experimental Impact Estimates in Effect Size Units, by Data Set and RD Estimation Approach

		Local	Difference[a]	95% Confidence
	RD	Experimental	(RD-RA)	Interval
RD High Sample				
Ed Tech				
Local Linear	-0.06	0.07**	-0.13	-0.27 to 0.02
	(0.12)	(0.03)	(0.07)	
Parametric	0.00	0.07**	-0.07	-0.23 to 0.08
	(0.05)	(0.03)	(0.08)	
TFA—Math				
Local Linear	0.05	0.12**	-0.08	-0.51 to 0.36
	(0.23)	(0.04)	(0.22)	
Parametric	0.07	0.12**	-0.06	-0.36 to 0.25
	(0.17)	(0.04)	(0.16)	
TFA—Reading				
Local Linear	0.05	0.02	0.03	-0.37 to 0.42
	(0.20)	(0.04)	(0.20)	
Parametric	0.03	0.02	0.01	-0.22 to 0.24
	(0.11)	(0.04)	(0.12)	
RD Low Sample				
Ed Tech				
Local Linear	0.08	0.07**	0.01	-0.15 to 0.16
	(0.05)	(0.03)	(0.08)	
Parametric	0.14**	0.07**	0.07	-0.07 to 0.21
	(0.05)	(0.03)	(0.07)	
TFA—Math				
Local Linear	0.07	0.12**	-0.05	-0.64 to 0.54
	(0.30)	(0.04)	(0.30)	
Parametric	0.22	0.12**	0.10	-0.15 to 0.35
	(0.14)	(0.04)	(0.13)	
TFA—Reading				
Local Linear	-0.03	0.02	-0.05	-0.42 to 0.32
	(0.19)	(0.04)	(0.19)	
Parametric	0.04	0.02	0.02	-0.23 to 0.27
	(0.14)	(0.04)	(0.13)	

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: RD estimates based on models presented in Chapters 3 and 4 and Appendix A. Experimental estimates based on estimates presented in Tables V.1 through V.3. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample. Restricted bandwidth for the local experimental estimate is the average of the optimal bandwidth from the RD High and RD Low samples.

[a] Standard errors calculated using 1000 bootstrap replications. For each replication, we select a new sample with replacement and calculate experimental, local linear and parametric estimates. The reported standard error is the standard deviation of the difference between the experimental and RD estimate over those 1000 samples.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Based on a test of the statistical significance of the difference between the RD and experimental impact estimates of these two educational interventions, the two approaches consistently resulted in impact estimates that were not significantly different from one another. Across the twelve comparisons of RD versus experimental impact estimates in the RD High and RD Low samples, there were no cases in which the difference was statistically significant at the 0.05 level. For example, the RD High analysis shows that for the Ed Tech intervention, the effect size difference between the local linear RD estimate and the experimental estimate was -0.13 effect size units, with an RD estimate of -0.06 and an experimental estimate of 0.07. This difference was not statistically significant, and the 95 percent confidence interval of the difference estimate ranged from -0.27 to +0.02. The difference between the parametric RD High and experimental Ed Tech impact estimates was -0.07 effect size units (0.00 versus 0.07, with a 95 percent confidence interval of the difference ranging from -0.23 to +0.08), and was also not statistically significant. The effect size differences for TFA were of a similar magnitude. Across the math and reading local linear and parametric estimates, the effect size differences ranged from -0.09 to 0.02. In each case, the 95 percent confidence intervals of the difference estimates included zero. For example, the estimated effect size of TFA on reading achievement was 0.05 based on the local linear RD design and 0.02 based on the experimental design; the 95 percent confidence interval of this 0.03 effect size difference was -0.37 to +0.42. [49] In the RD Low sample, the differences between the RD and experimental impact estimates ranged from -0.05 (for the local linear TFA math and reading impacts) to 0.10 (for the parametric TFA math impact).

Based on other criteria for assessing the comparability of the RD and experimental impact estimates, the evidence is mixed. We first examined whether the RD and experimental estimates had the same sign and general magnitude.[50] In the RD High sample, none of the Ed Tech impact estimates was very large, but the experimental estimate was small and positive, while the local linear RD estimate was negative and the parametric RD estimate was zero. For TFA, all of the RD and experimental impact estimates were positive, and both the local linear and parametric estimates of the impact on reading achievement were within 0.03 effect size units of the experimental estimate. In the case of math achievement, however, the experimental estimate was about twice as large as the RD estimates (0.12 versus 0.05 in the case of the local linear specification and 0.12 versus 0.07 in the case of the parametric specification). In the RD Low sample, all of the Ed Tech impact estimates were positive, but the parametric RD estimate was twice as large as the experimental estimate (0.14 versus 0.07). For TFA, the RD and experimental estimates are relatively close to one another

[49] While the differences between the RD and experimental effect sizes were of a similar magnitude based on the Ed Tech versus TFA data sets, the standard errors of both the RD and experimental estimates were approximately twice as large for the estimates based on TFA data than they were when based on Ed Tech data. This difference in standard errors reflects the smaller sample size of the TFA evaluation.

[50] Defining what we meant by estimates having "the same general magnitude" was necessarily arbitrary. One might consider two estimates as being dissimilar if they differed by more than 0.05 effect size units. For an elementary school sample, the average yearly achievement gain ranges from 0.40 to 1.52, depending on the grade and subject of the test; thus, an effect size of 0.05 corresponds to 3 to 12 percent of the estimated amount of learning in a year (Hill et al. 2007). An alternative approach would be to define as similar estimates within 0.10 effect size units of one another, since evaluations of education interventions are not typically powered to detect impacts smaller than this (in an examination of 40 evaluations of education interventions, Spybrook and Raudenbush [2009] found that minimum detectable effect sizes ranged from about 0.18 to 0.90).

(within 0.05 effect size units) in three of four cases. For TFA math, both the parametric RD and experimental estimates were positive, but the former was 0.10 effect size units larger.

We also examined whether the RD and experimental estimates had the same sign and level of statistical significance, which would be relevant for policy makers. In several cases, we found that while the RD and experimental impact had the same sign in most instances, they frequently differed in their level of statistical significance. Across the RD High and RD Low samples, in particular, there were five cases in which the experimental impact estimate was statistically significant and the RD estimate was not. In other words, a study based on this particular experimental design would have concluded that TFA had a positive impact on students' math achievement, while a study based on this particular RD design (using either a parametric or local linear specification) would not have found a statistically significant, positive impact of TFA on this outcome. It should be noted that differences in the statistical significance could arise either because of differences in the RD and experimental impact estimates themselves and/or differences in the statistical power of the two designs. In two of the five cases noted above, for example, the experimental estimate is statistically significant while the RD estimate is not, even though the magnitude of the RD estimate is about the same as or larger than the experimental estimate.

A final approach to comparing the RD and experimental estimates was to examine the full distribution of impact estimates from our bootstrap replicate samples. The bootstrap process provided 1,000 RD and experimental impact estimates, which we used to trace out the sampling distribution of the RD and experimental impact estimates. The sampling distributions of the effect size estimates based on the RD versus experimental design were not comparable, largely because the spread of the RD sampling distribution was larger than that of the experimental estimate. One indicator of the difference between the two distributions was that only a relatively small proportion of the RD distribution fell within the 95 percent confidence interval of the original experimental impact estimate.[51] In the case of Ed Tech, for example, only 52 percent of local linear RD impact estimates fell within the 95 percent confidence interval of the experimental impact estimate (Table V.5). For the parametric RD estimator, only 45 percent of this distribution fell within the 95 percent confidence interval of the original impact estimate. In the case of TFA math and reading, between 35 and 53 percent of the distribution of RD impact estimates fell within the 95 percent confidence interval for the analogous experimental estimate.

[51] If the RD and experimental estimators we examine here were perfectly comparable and produced the same sampling distributions, one would expect that approximately 95 percent of the RD distribution would fall within the 95 percent confidence interval for the experimental impact estimate.

Table V.5. Comparison of the Sampling Distributions of Regression Discontinuity (RD) Versus Experimental Impact Estimates, by Data Set and RD Estimation Approach

	Proportion of RD Impact Estimates within the 95% Confidence Interval of the Local Experimental Impact
Ed Tech	
Local Linear	0.52
Parametric	0.45
TFA—Math	
Local Linear	0.34
Parametric	0.39
TFA—Reading	
Local Linear	0.34
Parametric	0.53

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: RD estimates based on models presented in Chapters 3 and 4. Experimental estimates based on the local estimates of the treatment effect and its standard error presented in Tables V.1 through V.3. The sampling distributions were generated from 1000 bootstrap replications.

Although we found no statistically significant difference between the RD and experimental estimates of the impacts of Ed Tech and TFA, other tests of the comparability of the two sets of estimates suggest reason for concern. In particular, Table V.5 indicates that a relatively small proportion of the RD impact estimate sampling distribution falls within the 95 percent confidence interval of the local experimental impact estimate. Bias in the RD estimator could lead to a result like this. Even if there were no bias in the RD estimator, however, the limited statistical power of the individual RD estimates could lead to a sampling distribution with relatively little overlap with the experimental confidence interval. In other words, we could not estimate the impacts of these interventions precisely using an RD design (given our sample sizes).[52] For a given sample, the RD estimate may have differed from the experimental estimate by a substantively meaningful and policy relevant amount, even if the difference was not statistically significant. While the differences between the RD and experimental estimates were not statistically significant, we would have needed to observe a large difference between them to reject the null hypothesis that they were equal. The absolute values of the minimum estimated effect size differences that would have rejected the null hypothesis are reported in Table V.6. In four of the six comparisons, if we had found an effect size difference of 0.20, we would not have been able to reject the null. For each of the individual RD-experimental differences that we did observe, we could not distinguish between the possibility that the difference between the two estimation approaches arose by chance—that is, due to sampling variability—versus the possibility that there was a true difference between these estimation approaches.

[52] In the TFA analysis, the total sample used in the RD analysis was 804, but in the non-parametric RD case, only 518 observations were within the optimal bandwidth and used in the estimation. In the Ed Tech analysis, the total RD sample size was 3,681, with 1,513 observations falling within the optimal bandwidth.

Table V.6. Estimated Effect Size Difference Between the Regression Discontinuity (RD) and Experimental Impact Estimates, and Absolute Value of the Minimum Estimated Difference that Would Be Statistically Significant

	Actual Effect Size (ES) Difference	Absolute Value of the Minimum Estimated ES Difference that Would Reject the Null
Ed Tech		
Local Linear	-0.13	0.14
Parametric	-0.07	0.15
TFA—Math		
Local Linear	-0.08	0.44
Parametric	-0.06	0.31
TFA—Reading		
Local Linear	0.03	0.40
Parametric	0.01	0.23

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

C. Synthesizing Results

In the spirit of meta-analysis, we combined the results of six tests of the statistical significance of the difference between the RD and experimental impact estimates to summarize our evidence on the performance of the RD estimator. From the Ed Tech study, we used estimates based on both the RD High and an RD Low samples. From the TFA study, we used four sets of estimates—(1) the RD High sample for math, (2) the RD Low sample for math, (3) the RD High sample for reading, and (4) the RD Low sample for reading. Our overall estimate of the difference between an impact estimate based on an RD design versus one based on an experimental design was set to the mean value of the six individual difference estimates. The standard error of the overall difference was estimated using a bootstrap approach that accounted for any existing correlation between the estimates (such as correlation caused by the fact that for either the TFA RD High or RD Low sample the same individuals were used in estimating the RD-experimental difference for math and reading). We conducted this analysis separately for the comparison of the local linear RD versus local experimental impact estimate and the comparison of the parametric RD versus local experimental impact estimate.

Two key limitations of this analysis should be noted. Unlike a traditional meta-analysis, we are aggregating estimates across different interventions (Ed Tech and TFA) instead of different estimates of the impact of a single intervention. Our rationale for this is that we are interested not in the estimated impact of the intervention itself, but in how that estimate differs when it is based on an RD design versus and experimental design. Second, policy makers make decisions based on individual impact estimates, and so would be concerned about differences between RD and experimental estimates in each individual comparison, even if average differences across multiple comparisons were close to zero. In comparing the individual RD and experimental impact estimates in this study, we sometimes found cases in which the magnitude of the impact estimate and its level of statistical significance differed in policy-relevant ways. In these cases the impacts estimated using the RD design had a large standard error, highlighting the importance of obtaining adequate sample sizes in evaluations that use RD designs.

When we combined evidence from the six replication tests, we have stronger evidence regarding the estimated impact of a given intervention based on an RD design versus an experimental design.[53] The average effect size difference between the RD and experimental estimates was less than 0.05 (Table V.7). The average effect size difference for the parametric RD was 0.003, and the average effect size difference for the local linear RD was -0.02. Neither of these estimated differences was statistically significant. The second column shows the average absolute effect size difference, calculated by taking the absolute value of each difference before averaging. This measure tells us how far apart the RD and local experimental estimates are on average, without allowing positive and negative differences to offset one another. This absolute difference is a measure of the variability of the individual RD-experimental comparisons rather than a measure of the bias of the RD estimator. The absolute differences average 0.13 effect size units for the local linear RD and 0.09 effect size units for the parametric RD.

Table V.7. Estimated Difference Between Regression Discontinuity (RD) and Experimental Impact Estimates in Effect Size Units, Based on Aggregated Evidence from Six Replication Tests

	Average Effect Size (ES) Difference	Average Absolute ES Difference
Local Linear	-0.02 (0.08)	0.13
Parametric	0.003 (0.04)	0.09

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: The overall RD-experimental effect size difference is based on the estimates from the three RD High samples shown in Table V.6 as well as estimates from the three RD Low samples shown in Appendix A. The average and average absolute effect size differences were calculated for each of the 1000 bootstrap replications. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[53] This conclusion rests on the assumption that any biases in the RD design would be consistent across RD-experimental comparisons. In other words, we are assuming that the relatively small average RD-experimental difference across the six comparisons did not simply result from negative and positive RD biases cancelling each other out.

VI. SUMMARY OF RESULTS

In this report, we set out to examine regression discontinuity (RD) designs, and whether they produced similar impact estimates as experimental designs in two different studies of education interventions. We examined data from the Educational Technology (Ed Tech) Study and the Teach for America (TFA) Study. In each case, we generated experimental impact estimates based on the original experimental data and then used a subset of the original data chosen in such a way that it could have come from a well-implemented RD study of the same intervention. We then estimated impacts of each intervention using the RD approaches and compared these estimates to the experimental estimates. Through this comparison, we could assess the extent to which the RD design produced results similar to the gold-standard experimental results.

A key limitation to this approach for assessing the performance of the RD design is that the method we used for constructing the RD analysis sample artificially ensured that the design was well implemented. In other words, we could be certain that the design was based on a single, well-defined assignment variable, that assignment to the intervention depended solely (and perfectly) on whether the value of the assignment variable was above or below a clear cutoff value, and that there was no manipulation of values of the assignment variable by individuals for purposes of gaining access to the intervention. As a result, the main comparison of RD to experimental impact estimates (presented in Chapter V) does not allow us to fully assess the overall RD design, as we did not allow certain breakdowns in the implementation of the design. However, our RD-experimental comparison does allow us to assess the success of specific aspects of the RD approach—most importantly, the success of this design (in the context of two experimental studies) in modeling the relationship between the assignment variable and outcome and using the estimated discontinuity in this relationship at the cutoff value of the assignment variable as an estimate of the program impact.

We addressed one aspect of this limitation in an additional analysis presented in Appendix C. We altered the process for constructing the RD analysis samples so that we could mimic situations in which there was manipulation of the assignment variable on the part of individual students. In other words, we constructed alternative versions of the RD analysis sample so that some students whose true value of the assignment variable made them ineligible for the intervention (and hence should have put them in the control group) altered the value of this variable so that they ended up receiving the treatment and being in the treatment group. We then estimated impacts based on this manipulated RD analysis file, and compared them to impact estimates that would have resulted had no manipulation taken place. This analysis allowed us to examine the implications of different forms and different frequencies of manipulation on RD impact estimates, and examined whether a statistical test designed to detect manipulation of the assignment variable succeeded in doing so.

In this chapter, we summarize key results from our analysis. Section A covers the basic RD-experimental comparisons. Section B describes the results of the analysis of assignment variable manipulation.

A. Do RD Impact Estimates Match Those Based on an Experimental Design?

To compare the RD and experimental approaches, we used six separate tests that varied based on the study used, the outcome examined, and the way in which the RD sample was constructed. In particular, we made separate comparisons of estimates based on Ed Tech and TFA data, and in the case of TFA conducted separate analyses using reading and math test scores. In each case, we constructed two separate RD samples, which allowed two separate RD impact estimates. We labeled one such sample RD High, where students with pretest values above the median formed the treatment group and those with pretest values below the median formed the control group. The RD Low sample was based on the premise that the opposite procedure was used to determine assignment to treatment.

The statistical power of any one of these RD-experimental comparisons was relatively low, largely due to the limited statistical power of the RD design. However, the fact that we conducted six separate tests boosted the statistical power of the overall RD-experimental comparison. To take advantage of these repeated tests, we calculated the average difference between the RD and experimental impact estimate across the six tests.

Table VI.1 shows the key results from the six original RD-experimental comparisons along with their average. In each case, we assessed the performance of two versions of the RD model, a parametric specification and a non-parametric (local linear) specification. In the RD models, we used the optimally selected parametric and local linear specifications.

We found that the RD and experimental impact estimates were not statistically different from one another in any of the six comparisons, regardless of whether the comparison was based on the parametric or local linear RD specification. Among the twelve estimates of the RD-experimental difference, the values ranged from -0.09 to +0.10 in effect size or standard deviation units. The magnitudes of six of the estimates were positive, and the remaining six were negative.

These original comparisons did highlight one major issue, the limited statistical power of the RD impact estimates resulting in limited power of the RD-experimental comparison. While none of the estimates was significant, differences between the two approaches as large as 0.09 or 0.10 standard deviation units could be substantively important in practice. For example, in one case, the impact of the Ed Tech intervention was estimated to have an effect size of only 0.04 (and not statistically significant) based on the experimental design, but to have a statistically significant effect size of 0.14 based on the parametric RD model. We calculated the average RD-experimental difference across the six comparisons (separately for the parametric and local linear RD specifications). When we did so, we again found no statistically significant difference between the estimates. The magnitude of the RD-experimental difference was 0.02 effect size units in the case of the parametric RD model and -0.03 effect size units in the case of the local linear RD model.

B. How Does Manipulation of the Assignment Variable Influence RD Estimates?

In Appendix C, we explore the implications of relaxing the assumption of the perfect implementation of the assignment-to-treatment mechanism. In other words, we examine what might happen to the estimated impact based on the RD design if there is some manipulation of the assignment variable, or cheating.

Not surprisingly, the answers to these questions depend upon the circumstances surrounding the manipulation. Key dimensions that affect how cheating influences the RD design include the proportion who cheat, whether those who cheat come entirely from among students with pretest scores that almost—but not quite—make them eligible for the intervention, and whether those who cheat are those who have unobserved characteristics that would tend to cause them to do well (or poorly) regardless of the intervention. We found that the McCrary test successfully detected the presence of manipulation in some circumstances but not in others, and that the performance of the test was sensitive to the statistical power of the underlying RD design.

Table VI.1. Summary of Regression Discontinuity (RD) Versus Experimental Impact Estimates in Effect Size Units—by Data Set and Sample and RD Estimation Approach

		RD Impact Estimate		Difference Between RD & Experimental Estimate[b] (p-values in parentheses)	
Data and Sample	Local Experimental Estimate[a]	Parametric	Local Linear	Parametric RD	Local Linear RD
RD High Sample					
Ed Tech	0.07**	0.00	-0.06	-0.07 (0.08)	-0.13 (0.07)
TFA—Math	0.12**	0.07	0.05	-0.06 (0.16)	-0.08 (0.22)
TFA—Reading	0.02	0.03	0.05	0.01 (0.12)	0.03 (0.20)
RD Low Sample					
Ed Tech	0.07**	0.14**	0.08	0.07 (0.07)	0.01 (0.08)
TFA—Math	0.12**	0.22	0.07	0.10 (0.13)	-0.05 (0.30)
TFA—Reading	0.02	0.04	-0.03	0.02 (0.13)	-0.05 (0.19)
Average	---	---	---	0.003 (0.04)	-0.02 (0.08)

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: RD estimates based on models presented in Chapters 3 and 4. Experimental estimates based on estimates presented in Tables V.1 through V.3.

[a] Restricted bandwidth is the average of the optimal bandwidth from the RD High and RD Low samples.

[b] Standard errors calculated using 1000 bootstrap replications. For each replication, we select a new sample with replacement and calculate experimental, local linear and parametric estimates. The reported standard error is the standard deviation of the difference between the experimental and RD estimate over those 1000 samples.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

In examining different hypothetical instances of manipulation of the assignment variable, three clear findings about the influence of manipulation on the RD impact estimate emerged: 1) manipulation of the assignment variable had a substantial influence on the RD impact estimate when it was correlated with unobserved student characteristics that affected the ultimate outcome; 2) bias due to manipulation was larger when a larger proportion of control group students cheated; and 3) bias due to manipulation tended to be modest unless those who cheated came from among students with pretest scores close to the cutoff value. These patterns suggest specific issues to be aware of when considering a RD design to study effects of interventions that use test scores as the assignment variable. The results also highlight the importance of examining both institutional and statistical evidence of manipulation. And in addition to focusing only on whether manipulation of the assignment variable may be occurring in a RD design, researcher should also attempt to clarify the nature of the manipulation, since certain types of manipulation may be more problematic than others.

C. Limitations of the RD-Experimental Comparisons

The analysis and comparisons we have presented in this report shed light on the performance of RD models, but two key limitations must be kept in mind. First, conclusions presented here have been based on the analysis of data from just two studies. Conditions in these studies may differ from other potential applications of RD analysis, and so the performance of RD models could differ in these other contexts as well. Two dimensions on which other studies could have differed from the Ed Tech and TFA study in ways that could influence the results are worthy of note. RD designs could perform differently when examining interventions with very different impacts than the two studies examined here. And perhaps even more importantly, the performance of RD models could differ for evaluations in which the underlying relationship between the assignment variable and outcome was very different than the relationship between a pretest and posttest score that was the basis of the RD models examined in this report.

Second, the design upon which the RD estimates in this report have been based was not implemented under real world conditions. Instead, we created an analysis sample that could have been produced by an RD design, if the RD design was well implemented. The report's comparisons (especially those in the first five chapters) do not account for implementation problems encountered in some real world RD designs, although we did investigate the implications of manipulation of the assignment variable.

Future research can address each of these limitations. Using the basic approach we have modeled here, but with other experimental data sets, may shed light on the performance of RD models in different contexts. And replication studies in which both the experimental and RD designs were implemented under real world conditions—similar to replication studies conducted by Aiken et al. (1998), Skoufias (2003), and Black et al. (2007)—may shed light on the likelihood and implications of potential implementation problems in RD designs.

REFERENCES

Aiken, LS, SG West, D Schwalm, J Carroll, and S Hsuing (1998). "Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remedial Writing Program." *Evaluation Review* 22(4), 207-244.

Barker, L.E., E.T. Luman, M.M. McCauley, and S.Y. Chu (2002). "Assessing Equivalence: An Alternative to the Use of Difference Tests for Measuring Disparities in Vaccination Coverage." *American Journal of Epidemiology* 156, 1056-1061.

Black, D., J. Galdo, and J.A. Smith (2007). "Evaluating the Bias of the Regression Discontinuity Design Using Experimental Data." Working paper, January 2007.

Buddelmeyer, H. and E. Skoufias (2003). "An Evaluation of the Performance of Regression Discontinuity Design on PROGRESSA." Bonn, Germany: IZA.

Cameron, A.C. and P.K. Trivedi (2005). *MICROECONOMETRICS: Methods and Applications*. New York: Cambridge University Press, 2005.

Campuzano, L., Dynarski, M., Agodini, R., and Rall, K. (2009). "Effectiveness of Reading and Mathematics Software Products: Findings from Two Student Cohorts". Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2009-4041.

Cappelleri, J.C. and W.M.K. Trochim (1994). "An Illustrative Statistical Analysis of Cutoff-Based Randomized Clinical Trials." *Journal of Clinical Epidemiology* 47 (3), 261-270.

Cook, T.D. (2008). "Waiting for Life To Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics, and Econometrics." *Journal of Econometrics* 142 (2), 636-654.

Cook, T.D. and V.C. Wong (2008). "Empirical Tests of the Validity of the Regression Discontinuity Design." *Annales d'Economie et de Statistique* (91/92), 127-150.

Crump, R.K. & V.J. Hotz & G.W. Imbens & O.A. Mitnik (2008). "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics* 90(3), 389-405, 06.

Decker, P.T., D.P. Mayer, and S. Glazerman (2004). "The Effects of Teach for America on Students: Findings from a National Evaluation." Princeton, NJ: Mathematica Policy Research.

Dynarski, M., R. Agodini, S. Heaviside, T. Novak, N. Carey, L. Campuzano, B. Means, R. Murphy, W. Penuel, H. Javitz, D. Emery, and W. Sussex (2007). "Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort." Washington, DC: Institute for Education Sciences, National Center for Education Evaluation and Regional Assistance, NCEE 2007-4005.

Imbens, G. and K. Kalyanaraman (2009). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." National Bureau of Economic Research Working Paper, #14726.

Imbens, G. and T. Lemieux (2008). "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2), 615-635.

Hahn, J., P. Todd, and W. Van Der Klaauw (2001). "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design." *Econometrica* 69, 201-209.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey (2007). "Empirical Benchmarks for Interpreting Effect Sizes in Research." MDRC Working Paper on Research Methodology. New York, NY: MDRC.

Ludwig, J. and D. Miller (2005). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." National Bureau of Economic Research Working Paper, #11702.

McCrary, J. (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142 (2), 698-714.

Rogers, J., K. Howard, and J. Vessey (1993). "Using Significance Tests to Evaluate Equivalence between Two Experimental Groups." *Psychological Bulletin* 113(3), 553-565.

Schochet, P. (2008). "Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations." Washington, DC: Institute for Education Sciences, National Center for Education Evaluation and Regional Assistance, NCEE 2008-4026.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., Smith, J. (2010). Standards for Regression Discontinuity Designs. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf.

Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L.Bickman (Ed.), *Validity and social experimentation: Donald Campbell's legacy*. Thousand Oaks, CA: Sage, 13-35.

Spybrook, J. and S. Raudenbush (2009). "An Examination of the Precision and Technical Accuracy of the First Wave of Group Randomized Trials Funded by the Institute of Education Sciences." *Educational Evaluation and Policy Analysis* 31(3), 298-318.

Trochim, W.M.K and J.C. Cappelleri (1992). "Cutoff Assignment Strategies for Enhancing Randomized Clinical Trials." *Controlled Clinical Trials* 13, 190-212.

APPENDIX A. RD LOW ESTIMATION OF THE IMPACT OF ED TECH AND TFA

Our primary RD analysis assumed that students with pretest scores above the median received the treatment and students below the median were in the control group. To create this RD design, we discarded half of the data – treatment students with pretest scores below the median and control students with pretest scores above the median. The remaining (non-discarded) sample became the RD High sample. The discarded data, referred to as the RD Low sample, included treatment students with pretest scores below the median and control students with pretest scores above the median. We repeated the replication exercise with the RD Low sample for the estimated impacts of Ed Tech, TFA on math achievement, and TFA on reading achievement. This appendix reports the RD Low impact estimates and compares the RD Low impact estimates to the experimental results.

A. Ed Tech Regression Discontinuity Estimates

In the Ed Tech RD Low sample, students whose pretest scores were below the median received the treatment and students whose pretest scores were above the median did not. This assignment mechanism is reflected in the summary statistics presented in Table A.1. The treatment group had a mean pretest score of 34.0 points, while the control mean was 66.6 points. The differences in pretest and posttest scores were significant at the 0.01 level. While the treatment students in the RD Low sample had very different test scores from the treatment students in the RD High sample, the RD impact is still estimated for students at the median.

We estimated both nonparametric (local linear) and parametric RD models, using an optimal bandwidth selection procedure for the local linear models and a pre-specified procedure (described in Chapter II) for choosing the appropriate parametric specification. We followed the same algorithms we used for the RD High sample considering both graphical evidence and analytic models.

Table A.2 presents the estimated impacts of the Ed Tech intervention from parametric models and the RD Low sample. The first column shows the regression results for the linear parametric specification. The specification in the second column allows for a quadratic relationship between the assignment variable and the outcome, and the third column allows for a cubic relationships. The F-test on the added terms indicates that the quadratic specification is the optimal parametric RD model. The quadratic specification was also optimal for the Ed Tech RD High sample. In the quadratic specification, the estimated impact of Ed Tech was 2.82 points, and the impact was statistically significant. The estimated impact was sensitive to the parametric specification. With a linear model, the estimated impact was -0.57, and not statistically significant. In the cubic specification, the estimated impact was 2.33, and not statistically significant.

In the local linear model, we used the Imbens-Kalyanaraman procedure to select the optimal bandwidth for the RD Low sample. For the Ed Tech Low sample, the optimal bandwidth was 14.3. Table A.3 presents the estimated impacts of the Ed Tech intervention using the nonparametric local linear specification with the RD Low sample. Specification 1, using the optimal bandwidth, yielded an impact estimate of 1.57, which was not statistically significant. As alternative specifications, we estimated local linear models with bandwidths half as large and twice as large as the optimal bandwidth. With both of these alternative bandwidths, the estimated impact of Ed Tech was positive and statistically significant.

Table A.1. Characteristics of the Ed Tech Rd Low Sample

	Full Sample		Control		Treatment	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Student Characteristics						
Proportion Female	0.50	0.50	0.53	0.50	0.47**	0.50
Age (Years)	9.51	2.10	9.55	2.11	9.46	2.10
Proportion In Treatment Classroom	0.51	0.50	0.00	0.00	1.00	0.00
Pretest Score[a]	50.09	20.10	66.62	12.66	33.98**	10.74
Re-centered Pretest Score[b]	0.08	20.10	16.62	12.66	-16.01**	10.74
Posttest Score[a]	47.88	19.87	60.51	16.06	35.59**	14.89
Re-centered Posttest Score[b]	-0.24	19.33	12.32	15.43	-12.41**	14.26
Teacher Characteristics						
Proportion Female	0.87	0.33	0.90	0.30	0.85	0.36
Years of Teaching Experience	10.60	9.13	10.66	9.40	10.56	9.90
Sample Size						
Students	**3888**		**1684**		**2204**	
Teachers	**345**		**149**		**196**	

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original Ed Tech study.

[a] Test scores are reported in NCE units. The test has a mean of 50 and SD of 21.06.

[b] Re-centered test scores are calculated by subtracting the grade-specific median from the original test score. This results in an RD cutoff score of zero for all grades.

* Difference between treatment students and control students is significantly different from zero at the .05 level, two-tailed test.

** Difference between treatment students and control students is significantly different from zero at the .01 level, two-tailed test.

Table A.2. Estimated Impact of Treatment Status on Test Scores, Regression Discontinuity Parametric Specifications—Ed Tech

	Specification 1	Specification 2	Specification 3
RD Low Model	Linear	Quadratic	Cubic
Treatment Status	-0.57 (0.75)	2.82** (0.96)	2.33 (1.20)
Pretest	0.77** (0.02)	1.14** (0.63)	0.95** (0.14)
Pretest * Treatment	-0.04 (0.03)	-0.26** (0.10)	-0.02 (0.22)
Pretest Squared		-0.01** (0.00)	0.00 (0.00)
Pretest Squared * Treatment		0.01** (0.00)	0.00 (0.01)
Pretest Cubed			-0.00 (0.00)
Pretest Cubed * Treatment			0.00 (0.00)
Sample Size	**3888**	**3888**	**3888**
R-Squared	**0.67**	**0.67**	**0.67**
F-Test P-Value for Squared Terms		**0.00**	
F-Test P-Value for Cubed Terms			**0.28**

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also include random assignment block fixed effects and a teacher random effect. Specification 2 is the preferred model based on the F-tests presented at the bottom of the table.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Table A.3. Estimated Impact of Treatment Status on Test Scores, Regression Discontinuity Nonparametric Specifications—Ed Tech

	Specification 1	Specification 2	Specification 3
RD Low Model	Optimal Bandwidth	(1/2)*Optimal Bandwidth	2*Optimal Bandwidth
Treatment Status	1.57 (1.06)	3.23* (1.53)	1.71* (0.83)
Pretest	0.83** (0.08)	1.21** (0.24)	0.94** (0.04)
Pretest * Treatment	-0.04 (0.12)	-0.30 (0.35)	-0.17** (0.05)
Sample Size	**1904**	**1084**	**3270**
R-Squared	**0.32**	**0.23**	**0.57**
Bandwidth Size	**14.34**	**7.17**	**28.69**

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also include random assignment block fixed effects and a teacher random effect.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

B. TFA Regression Discontinuity Estimates

The TFA Math RD Low sample was created using the same assignment mechanism: students whose math pretest scores were below the median received the treatment and students whose math pretest scores were above the median did not. This assignment mechanism is reflected in the summary statistics presented in Table A.4. The treatment group had a mean math pretest score of 20.7 points, while the control mean was 44.6 points. The differences in math pretest and posttest scores were significant at the 0.01 level. Although reading scores were not explicitly part of the Math RD Low assignment mechanism, the differences in reading pretest and posttest scores were also statistically significant. Along with test score differences, there were also statistically significant demographic differences between treatment and control students in the RD Low sample. Treatment students were more likely to be black (72 percent compared to 56 percent) and less likely to be Hispanic (26 percent compared to 32 percent). Treatment students were also more likely to receive free lunch (83 percent compared to 78 percent).

The RD parametric impact estimates are presented in Table A.5. The F-test on the quadratic and cubic terms indicated that the optimal parametric specification was the quadratic. In the quadratic specification, the impact of TFA on math scores was 4.33, and statistically significant. In the alternative linear and cubic specifications, the estimated impacts were also positive, but not statistically significant.

For the nonparametric impact estimates presented in Table A.6, the optimal bandwidth for the TFA Math RD Low sample was 9. With this relatively small bandwidth, the nonparametric specification sample included 47 percent of the observations included in the RD Low parametric specifications. The estimated impact of TFA on math scores was 1.36, and not statistically significant. With a bandwidth half as large, the estimated impact was 4.87, and not statistically significant. With a bandwidth twice as large, the estimated impact was 4.00, and statistically significant.

Using discarded data from the TFA Reading High sample, we also created a TFA Reading RD Low sample. In the Reading RD Low sample, treatment students had reading pretest scores below the median, and control students had reading pretest scores above the median. For the parametric RD impact estimates, the optimal specification was the linear specification in the first column (see Table A.7). With a linear relationship between the reading pretest and posttest, the estimated impact of TFA on reading test scores was 0.75 points, and not statistically significant. Table A.7 also reports impact estimates from quadratic and cubic specifications. Neither estimate was statistically significant.

The nonparametric impacts estimates for the Reading RD Low sample are presented in Table A.8. The optimal bandwidth was 12.1 points. Using the optimal bandwidth, the nonparametric impact estimate was -0.60 and not significant. Table A.8 also includes impact estimates using smaller and larger bandwidths. Neither of these estimates was statistically significant. Regardless of the specification, all of the Reading RD Low parametric and nonparametric estimates of the TFA impact on reading test scores were not statistically significant.

Table A.4. Characteristics of the TFA Math RD Low Sample

	Full Sample		Control		Treatment	
Student Characteristics	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Proportion Female	0.49	0.50	0.50	0.50	0.48	0.50
Proportion Black	0.65	0.48	0.56	0.50	0.72**	0.45
Proportion Hispanic	0.29	0.45	0.32	0.47	0.26*	0.44
Age (Years)	8.5	1.6	8.5	1.5	8.4	1.6
Proportion Free Lunch	0.81	0.39	0.78	0.42	0.83*	0.37
Pretest Math Score[a]	31.8	18.2	44.6	14.8	20.7**	12.8
Pretest Reading Score[a]	29.8	19.0	37.6	18.8	22.9**	16.3
Posttest Math Score[a]	33.4	19.2	40.7	18.4	27.2**	17.7
Posttest Reading Score[a]	31.4	20.6	38.0	19.8	25.8**	19.6
Sample Size						
Students	**835**		**431**		**404**	
Teachers	**100**		**56**		**44**	

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original TFA study.

[a] Test scores are reported in NCE units. The test has a mean of 50 and SD of 21.

* Difference between treatment students and control students is significantly different from zero at the .05 level, two-tailed test.

** Difference between treatment students and control students is significantly different from zero at the 01 level, two-tailed test.

Table A.5. Estimated Impact of Treatment Status on Math Test Scores, Regression Discontinuity Parametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD Low Model	Linear	Quadratic	Cubic
Treatment Status	1.31 (1.58)	4.33* (2.00)	4.22 (2.52)
Pretest	0.77** (0.07)	1.10** (0.21)	1.31** (0.47)
Pretest * Treatment	-0.39** (0.10)	-0.43 (0.26)	-0.90 (0.56)
Pretest Squared		-0.01 (0.01)	-0.02 (0.03)
Pretest Squared * Treatment		0.02* (0.01)	0.02 (0.03)
Pretest Cubed			0.00 (0.00)
Pretest Cubed * Treatment			-0.00 (0.00)
Sample Size	**838**	**838**	**838**
R-squared	**0.35**	**0.35**	**0.36**
F-Test P-Value for Squared Terms		**0.03**	
F-Test P-Value for Cubed Terms			**0.51**

Source: Data from the Teach For America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect. Specification 2 is the preferred model based on the F-tests presented at the bottom of the table.

*Significantly different from zero at the .05 level, two-tailed test.

**Significantly different from zero at the .01 level, two-tailed test.

Table A.6. Estimated Impact of Treatment Status on Math Test Scores, Regression Discontinuity Nonparametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD Low Model	Optimal Bandwidth	(1/2)*Optimal Bandwidth	2*Optimal Bandwidth
Treatment status	1.36 (2.59)	4.87 (5.23)	4.00* (2.04)
Pretest	0.73 (0.40)	2.33 (1.84)	1.00** (0.16)
Pretest * Treatment	-0.62 (0.50)	-1.79 (2.06)	-0.47* (0.21)
Sample Size	**392**	**196**	**624**
R-Squared	**0.15**	**0.23**	**0.23**
Bandwidth Size	**9.0**	**4.5**	**18.0**

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Table A.7. Estimated Impact of Treatment Status on Reading Test Scores, Regression Discontinuity Parametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD Low Model	Linear	Quadratic	Cubic
Treatment Status	0.75 (1.68)	0.05 (2.23)	-2.89 (2.93)
Pretest	0.77** (0.07)	0.59** (0.21)	-0.40 (0.51)
Pretest * Treatment	-0.36** (0.10)	-0.08 (0.31)	1.32 (0.74)
Pretest Squared		0.00 (0.01)	0.06* (0.03)
Pretest Squared * Treatment		-0.00 (0.01)	-0.03 (0.05)
Pretest Cubed			-0.00 (0.00)
Pretest Cubed * Treatment			0.00 (0.00)
Sample Size	**797**	**797**	**797**
R-Squared	**0.45**	**0.45**	**0.45**
F-Test P-Value for Squared Terms		**0.58**	
F-Test P-Value for Cubed Terms			**0.07**

Source: Data from the Teach For America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect. Specification 1 is the preferred model based on the F-tests presented at the bottom of the table.

*/** Significantly different from zero at the .05/.01 level, two-tailed test.

Table A.8. Estimated Impact of Treatment Status on Reading Test Scores, Regression Discontinuity Nonparametric Specifications—TFA

	Specification 1	Specification 2	Specification 3
RD Low Model	Optimal Bandwidth	(1/2)*Optimal Bandwidth	2*Optimal Bandwidth
Treatment Status	-0.60 (2.67)	-6.53 (5.44)	0.45 (2.01)
Pretest	0.47 (0.32)	-1.09 (1.36)	0.72** (0.12)
Pretest * Treatment	0.16 (0.41)	0.93 (1.49)	-0.33* (0.15)
Sample Size	**408**	**222**	**682**
R-Squared	**0.28**	**0.27**	**0.34**
Bandwidth Size	**12.1**	**6.1**	**24.2**
Treatment Status	-0.60 (2.67)	-6.53 (5.44)	0.45 (2.01)

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports results of a regression model that also included an indicator for whether the pretest was taken in Spanish, random assignment block fixed effects, and a classroom random effect.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

C. Comparing the RD and Experimental Estimates

The previous two sections presented estimates of the impacts of the Ed Tech and TFA interventions using the RD Low sample. To assess the performance of these RD estimates, this section compares the RD and experimental estimates. The experimental benchmark for the RD Low impact estimates is the same experimental benchmark used for the RD High sample. The experimental results are reported in Chapter 5 (Tables V.1 – V.3).

Table A.9 reports the RD Low impact estimates, the experimental impact estimates, and the difference between the two estimates. To ease comparisons across the Ed Tech and the TFA studies, all of the impact estimates are reported in effect size units. For the Ed Tech study, the difference between the RD Low and experimental impact estimates was 0.04 for the local linear specification and 0.10 for the parametric specification. Neither difference was significantly different from zero. For TFA Math, the difference between the impact estimates was -0.06 for the nonparametric RD and 0.09 for the parametric RD. Again, we cannot reject the null hypothesis of no difference between the estimates. The differences between the TFA Reading impact estimates were also not statistically significant. The effect size difference for the nonparametric RD was -0.06 and the effect size difference for the parametric RD was 0.01.

As with the RD High sample, the differences between the RD and experimental impact estimates were not statistically different from zero. Chapter 5 discusses the main limitation of this finding, which is that the statistical power of the RD analysis—and the resulting comparison of RD and experimental estimates—was relatively low. To address this limitation, Chapter 5 reports average results based on evidence from estimated combined across the data sets and across the RD High and RD Low approaches.

Table A.9. Regression Discontinuity (RD) Versus Experimental Impact Estimates in Effect Size Units, by Data Set and RD Estimation Approach

	RD Low	Local Experimental	Difference[a] (RD-RA)	95% Confidence Interval
Ed Tech				
Local Linear	0.08 (0.05)	0.07** (0.03)	0.01 (0.08)	-0.14 – 0.16
Parametric	0.14** (0.05)	0.07** (0.03)	0.07 (0.07)	-0.07 – 0.21
TFA – Math				
Local Linear	0.07 (0.30)	0.12** (0.04)	-0.05 (0.30)	-0.64 – 0.53
Parametric	0.22 (0.14)	0.12** (0.04)	0.10 (0.13)	-0.16 – 0.36
TFA – Reading				
Local Linear	-0.03 (0.19)	0.02 (0.04)	-0.05 (0.19)	-0.42 – 0.32
Parametric	0.04 (0.14)	0.02 (0.04)	0.02 (0.13)	-0.24 – 0.27

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: RD estimates based on models presented earlier in this chapter. Original experimental estimates based on estimates presented in Tables V.1 through V.3. Restricted bandwidth for the local experimental estimate is the average of the optimal bandwidth from the RD High and RD Low samples.

[a]Standard errors calculated using 1000 bootstrap replications. For each replication, we select a new sample with replacement and calculate experimental, local linear and parametric estimates. The reported standard error is the standard deviation of the difference between the experimental and RD estimate over those 1000 samples.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

APPENDIX B. EXPLORING HETEROGENEITY IN EXPERIMENTAL IMPACT ESTIMATES

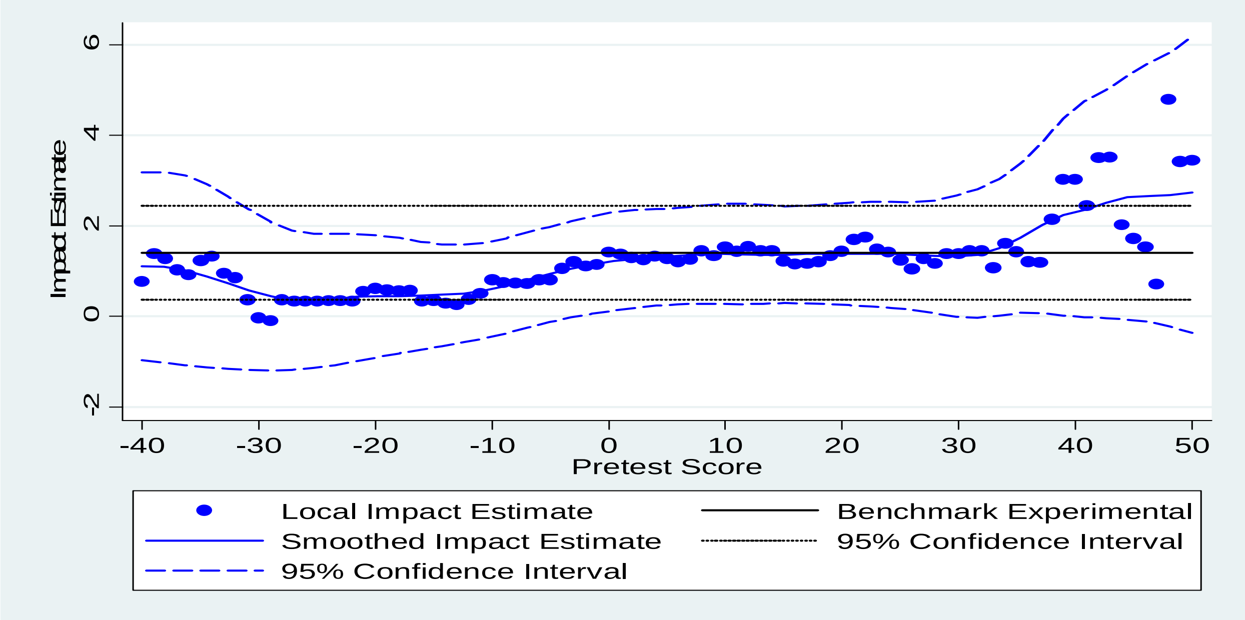
This appendix describes our additional investigation of treatment effect heterogeneity. As discussed in the main text, a key detail in the replication exercise is to choose the appropriate impact estimate to use as our true impact to be compared with the RD estimates. As described in Chapter II, we use the local experimental estimate centered around the cutoff, as our benchmark estimate for our main analysis. The analysis presented in this appendix is a supplementary analysis in the spirit of Crump et al. (2008) that explores the degree to which the experimental impact estimate varies over the range of pretest scores.

To produce the different impact estimates over different ranges of pretest scores we ran a series of regressions, each using the same model and bandwidth used for our main impact analysis presented in Chapter V. For each data set, we ran one regression centered at each point in the range of pretest scores, using all observations within the optimal bandwidth on either side of that point. For the Ed Tech data, the optimal bandwidth was 11.48 and the lowest recentered pretest score was -51, so the first regression was centered at -40 and used the subset of observations with pretest scores from -51.00 through -28.52. The second regression was centered at -39 and used the subset of observations with pretest scores from -50.48 to -27.52, and so on through the full range of data. We performed the same procedure with the TFA Math and TFA Reading data sets. The results of these regressions are presented in the figures below.

Figure B.1 presents the range of estimates for the Ed Tech data. The solid horizontal line represents the value of the benchmark experimental estimate from the main analysis (the local experimental estimate), and the dotted lines above and below the solid line represents the upper and lower bounds of the 95% confidence interval of this estimate. Each diamond is the impact estimate for the regression centered at the corresponding pretest score shown on the x-axis (in other words, the local impact estimate for that pretest score). The smoothed local polynomial approximation for the series of local impact estimates and the corresponding 95% confidence interval for the series of estimates are also included in this graph. These figures exclude the results from regressions that used fewer than 300 observations.[54] Figures B.2 and B.3 present the same information for TFA Math and TFA Reading, respectively. Note that in each figure the confidence bands get wider at extreme pretest values—this corresponds to smaller sample sizes for these values.

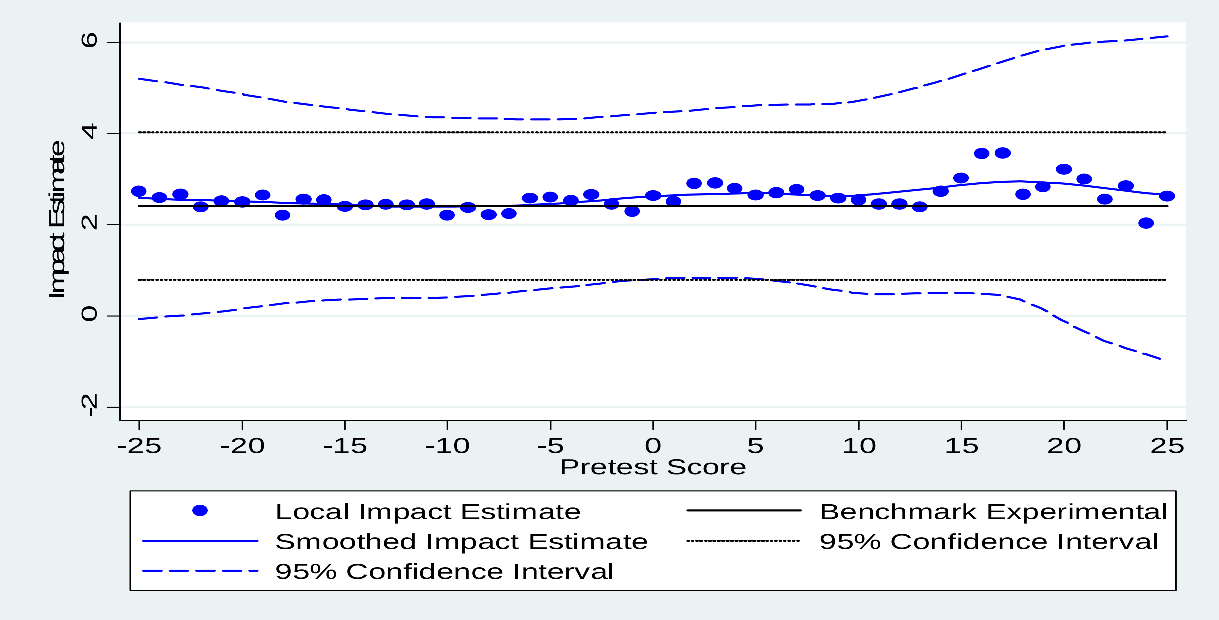
[54] This excludes a few points from extremes of the pretest distribution for the TFA data sets.

Figure B.1. Heterogeneity of RA Impacts – Ed Tech



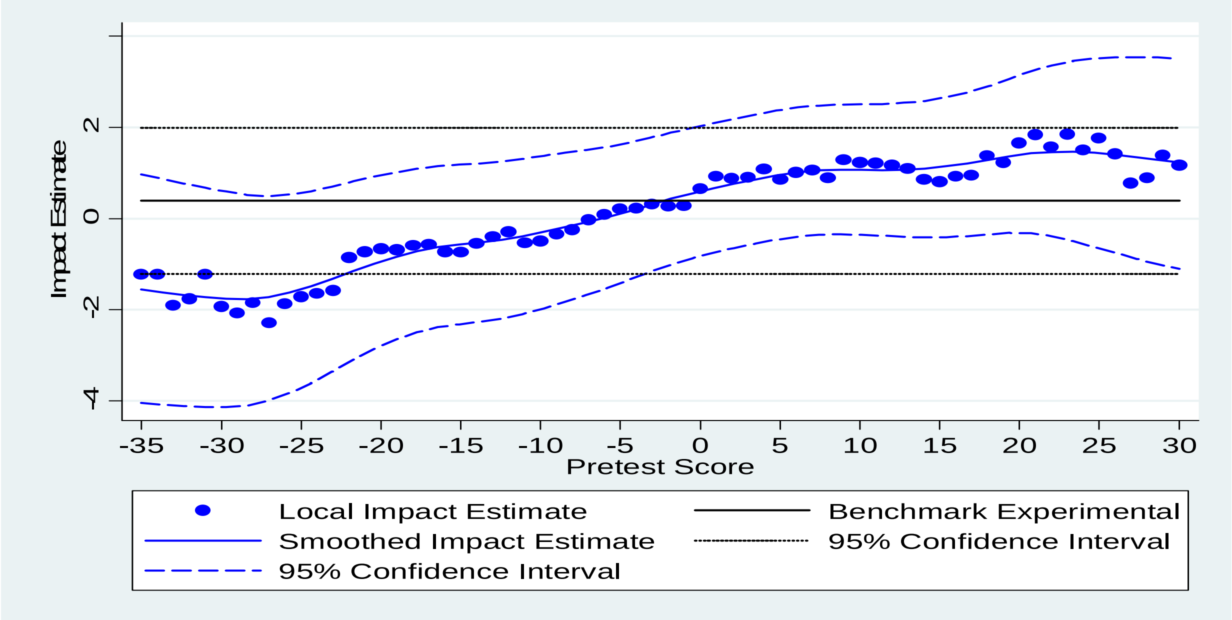
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure B.2. Heterogeneity of RA Impacts – TFA Math



Source: Data from the Teach for America Study (Decker et al. 2004).

Figure B.3. Heterogeneity of RA Impacts – TFA Reading



Source: Data from the Teach for America Study (Decker et al. 2004).

We arrive at two main conclusions from these figures. First, the estimated local impact estimate fell within the confidence interval for the benchmark estimate for the vast majority of cases across the range of pretest scores. Second, although in the case of TFA reading there did seem to be a pattern in the local impact estimates with these estimates getting more positive at higher pretest values, the local estimate at the pretest median—our cutoff value—was very close to the benchmark local experimental estimate. Because there is variation in the impact estimates across the range of pretest scores, it is possible that our results would have been very different if we used a cutoff point away from the median of the distribution.

APPENDIX C. MANIPULATION OF THE ASSIGNMENT VARIABLE

An RD design for estimating the impacts of an intervention is based on a situation in which assignment to the intervention is determined entirely by the value of a single characteristic. In our case, the assignment variable was the student pretest score. A key assumption of RD models is that the value of the assignment variable is exogenous; in particular, that it is not manipulated by program applicants or operators for purposes of receiving the treatment or avoiding the treatment. If there is manipulation of the assignment variable, then the assumption that treatment status is related only to the value of the assignment variable will likely be violated.

Since we constructed the RD design in this study, we were confident that manipulation was not a concern in the main analyses. However, it is possible to construct data sets based on an RD design in such a way that manipulation of the assignment variable has occurred. In this appendix, we examine the implications of manipulation of the assignment variable on RD estimates of impacts. While manipulation violates a key assumption of the RD design, it is not clear whether there is some level of manipulation that would not substantively influence the impact estimates, or whether any manipulation of the assignment variable invalidates the RD approach. Additionally, the influence of manipulation may depend on which students have manipulated test scores. A priori, a researcher would be more concerned about bias in situations where many scores were manipulated or where manipulation is correlated with anticipated gains. And while there is a specification test designed to detect manipulation (McCrary 2008), the relationship between the test's ability to detect the problem and the influence of manipulation on estimates is not clear.

For this analysis, we generated samples that could have resulted from RD designs in which some sample members altered their true value of the pretest variable in order to ensure that they received the treatment. In these samples, we varied the extent of the manipulation and the types of students with manipulated test scores. We then estimated impacts with these altered samples, as well as with the original, perfectly implemented RD sample, and measured how the impact estimates differed. We focused on the resulting bias in the RD impact estimate due to manipulation—defined as the difference between the estimate under the scenario with manipulation of the assignment variable and the estimate under the scenario with no manipulation. In addition, we conducted the McCrary test on the RD models estimated with the "manipulated" samples to see if the test detected the manipulation. For all analyses of manipulation, we used simulated data based on the original Ed Tech and TFA data, to generate the samples we would use to compare the performance of the RD estimator under various alternative scenarios of manipulation. Appendix D describes the approach we used to simulate additional data. We start with a simple case. In the RD High samples, individuals with pretest scores above the median received the treatment; those with scores below the median did not. To introduce manipulation to this assignment rule, we mimicked a scenario where some subset of individuals with pretest scores below the cutoff altered their test score to a value above the cutoff so they would receive the treatment. This sort of manipulation assumes that the treatment was desirable - a set of individuals went to some effort to ensure that they would receive the intervention. The manipulation, or cheating, could be caused by individual students if they have access to the scores or if the program depends on self-report of some information, or by teachers or program operators altering test results to ensure that certain students have access to the treatment.

A. Creating Data Set with Manipulation of Assignment Variable

Our empirical strategy involved comparing RD impact estimates using our original (or baseline) data set that had no manipulation of the assignment variable with the RD impact estimates based on several different versions of a data set that included some form of manipulation of the assignment variable. The data set we used in this analysis was based on our original data from the experimental analysis, supplemented with data simulated as described in Appendix D. The simulated data were added to increase the population of students near the RD cutoff values, which we then used in additional simulations of various manipulation scenarios. In particular, we created several different data sets using different values for the parameters representing the extent and type of manipulation. In this way, we examined the implications of different manipulation scenarios on both bias in the RD impact estimate and whether the McCrary test successfully identified the presence of manipulation.

We chose cases whose pretest values were to be manipulated at random from the pool of control observations, but we varied the likelihood of being chosen based on three key factors. For each factor, we examined a range of values to assess the effects of different levels of manipulation on the coefficient estimates.

1. How much of the original sample "cheats"? We varied the fraction cheating to allow from 1 percent to 10 percent of the sample members who would not otherwise have received the treatment to change their pretest values so that they did receive the treatment. We expect that as the fraction cheating increases the RD bias due to manipulation will also increase, holding the other parameters constant. We also expect that the McCrary test would be more likely to detect manipulation as the fraction cheating increases.
2. Are cheaters chosen from the whole distribution of control observations below the pretest cutoff, or are those with pretest values right below the cutoff more likely to cheat? We chose observations from within various ranges of pretest values. We began with a window of pretest values extending from the cutoff to 5 points below the cutoff, and then increased the window in stages by 5 points until we reached a window extending between the cutoff and 50 points below the cutoff. A small manipulation window would correspond to a situation where only students with pretest scores just below the cutoff manipulated their pretest value in order to receive the treatment, or if a teacher changed pretest values for a handful of students who scored right below the cutoff.

Is cheating correlated with an individual's posttest score holding constant their pretest score? We examined two variants of this parameter. First we assumed that among those with a given pretest value, those with the highest posttest values were most likely to change their pretest value to receive the treatment.[55] In other words, under this variant those with higher values of the error term in the outcome equation were most likely to manipulate their value of the assignment variable. This would be the case if highly motivated students who fell below the pretest cutoff were manipulating their

[55] In practice, for a given pretest value within the window selected for cheating, we randomly selected cases with posttest values above the conditional mean among those with that pretest value. Thus, cases with that pretest value and whose posttest value was below the conditional mean had no chance of being selected as a cheater under this variant.

score to receive the treatment or if teachers manipulated scores to gain access to the treatment for students whose performance on the pretest seemed unusually poor given their usual academic performance. Under the second variant, we relaxed this assumption and chose randomly from all available observations within the specified pretest window.

These three factors combined to form 200 possible manipulation scenarios for each data set. Under each of these scenarios, we used a four step process to go from the baseline situation in which there was no pretest manipulation, or cheating, to one of the scenarios with cheating. These steps are described below:

Step One: We first determined the number and type of observations to be manipulated under a given manipulation scenario, based on the fraction cheating and the specifics of the students whose scores were to be manipulated in that scenario. In particular, we specified which students would be eligible to cheat, based on their position in the distribution of pretest and posttest scores among the full set of control students in the baseline RD data set.

Step Two: We randomly selected these cases from the pool of all control students in the baseline data set based on the parameters of the manipulation scenario determined in Step One. These selected cases were the designated cheaters, but we could not use the exact values of their outcome (the posttest score) in the manipulated data set because these control students did not actually receive the treatment and so we did not have information for what their outcome would have been under the treatment.

Step Three: We dropped the selected control students from the data set and replaced them with matched observations from among the treatment students who had been discarded in the original analysis. We looked for treatment students in the discarded below-cutoff data that had the same pretest value as the control students selected to be manipulated.[56] Students were also matched on their grade level. In some of the manipulation scenarios, we also ensured that if the control student selected for manipulation was drawn only from among the top half of the control group posttest distribution (conditional on that pretest value), then the matched treatment student would be selected from among the top half of the treatment group posttest distribution (conditional on that same pretest value).

Step Four: We generated a new, manipulated pretest value above the cutoff for the matched treatment students selected in Step Three. For each case, we used a pretest score randomly selected from a window extending from the cutoff value to ten points above the cutoff value. The resulting data set has the same number of observations as the baseline data set, but will have more treatment observations. The treatment students who were added represent students who should have been in the control group and not received the treatment intervention, but who ended up manipulating their pretest score to a value above the cutoff value so that they would receive the intervention.

With this new simulated RD data set with manipulation, we could again estimate impacts using a basic RD model. The treatment group in this model included all the treatment students from the baseline RD model plus the additional cases with manipulated pretest values. The control group

[56] We first attempted to match students on the basis of having the same exact pretest value. If there was no match, we selected a treatment student whose pretest value was within an increasing number of centiles (up to four) of the pretest value of the control student selected for manipulation.

included the same control students from the baseline RD model except those that were dropped because they were selected to be cheaters. The new data set looked like a sharp RD data set, since all treatments appeared to have pretest scores above the cutoff and all controls had scores below the cutoff. Thus, we used sharp RD methods with the newly created data set to estimate impacts. We then compared this impact estimate with the RD impact estimate found using the non-manipulated data, with the difference between the two labeled the RD bias due to manipulation. For each manipulation scenario we also conducted the McCrary test for the presence of manipulation.[57] This test examines the continuity of the density of the assignment variable around the cutoff; any lumpiness in the density may suggest that values of the assignment variable were altered to affect treatment status.

B. Comparing RD Impact Estimates with Manipulation to Benchmarks

To illustrate the analysis, Table C.1 presents the Ed Tech estimates for two sets of manipulation parameters, each of which is based on a manipulation scenario in which 3 percent of the baseline control group manipulated their pretest score to gain the treatment, and cheaters were drawn from a window of 15 points below the cutoff.[58] To make this example more concrete, imagine a program that could serve only about half of a potential pool of 1,000 applicants. If such a program used a cutoff value of the test score median from the previous year, 3 percent cheating within a 15 point window would mean that 15 students were manipulating their score and that they would all be drawn from the students with scores within 15 points below the median. This would be about 6% of the students that fall within this window, assuming a normal distribution similar to the one in the standardized test scores we use.

In one of the two manipulation scenarios presented in the table, cheating was correlated with the posttest outcome (column 2), while in the other scenario it was not (column 3). Column 1 presents the benchmark estimates produced from the baseline (no-manipulation) data set. For the baseline case, the z-statistic for the McCrary test, which examined whether the density of the assignment variable just below the cutoff was significantly different from the density just above the cutoff, was -0.30 and not statistically significant. Figure C.1 shows the density graph that corresponds to this McCrary test. Each circle represents the density of observations falling within that cell of the assignment variable. The dotted and solid lines are local polynomial approximations of the density on each side of the cutoff point. For the baseline data set, the densities on each side of the cutoff are similar, with the local polynomial approximations nearly meeting at the cutoff. Thus, in the baseline case there was no evidence of cheating, which was not surprising given that we constructed the baseline RD data set to ensure that there would be no manipulation.

[57] See McCrary (2008) for more detail on this test. Note that the McCrary test only detects net manipulation across the treatment threshold. If some observations were being manipulated to receive treatment and some to avoid treatment, the local density might be unaffected.

[58] This type of manipulation mimics the situation where the cutoff is fixed and is not based on the distribution of the current population. Using our median rule, this could happen if the cutoff was based on the median test score from a previous year of data, so individuals seeking access to the treatment this year would know the cutoff and how their score compared to it.

Table C.1. Regression Results Using Manipulated Data (Ed Tech)

	Baseline Model[1]	With Manipulation	
	(1)	(2)	(3)
Fraction Cheating	0%	3%	3%
Manipulation Window	n/a	15 points	15 points
Is cheating correlated with outcome?	n/a	Yes	No
McCrary Test Z-stat	-0.30	1.86*	1.56
Local Linear Estimates			
Optimal Bandwidth	11.09	11.17	10.77
Impact Estimate	-0.14	0.30	-0.24
	(0.32)	(0.32)	(0.32)
Sample Size	19251	19386	19045
Bias due to manipulation (ES)	n/a	0.02	0.00
Parametric Estimates			
Impact Estimate	-0.07	0.37	-0.27
	(0.32)	(0.32)	(0.32)
Sample Size	46118	45568	45568
Bias due to manipulation (ES)	n/a	0.02	-0.01

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

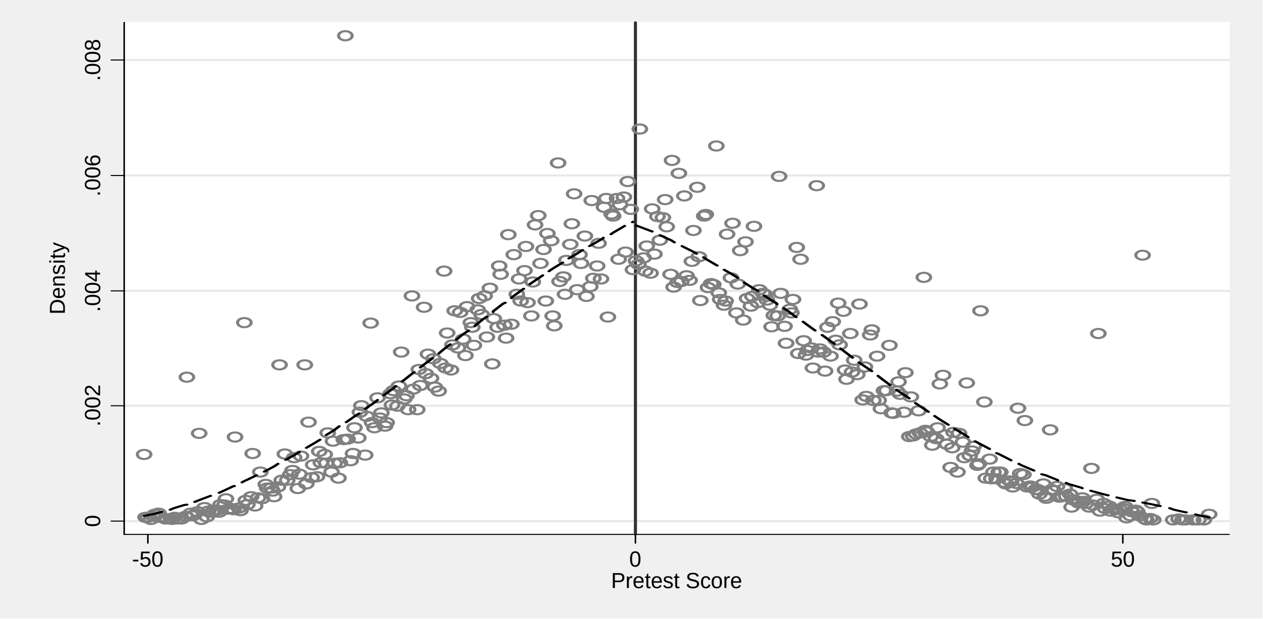
Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports selected coefficients from regression models that also included random assignment block fixed effects and a teacher random effect. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[1] The baseline model is the preferred RD model for the non-parametric or parametric specification using simulated data.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Figure C.2 and column 2 of Table C.1 present the results for the manipulation scenario in which cheating was correlated with the posttest value. In the figure, there is a gap between the density on the right and the left of the cutoff, but the McCrary test z-statistic of 1.86 indicated that this discontinuity was not statistically significant at the .05 level. The estimated impact of Ed Tech based on this manipulated data set was 0.30 NCE points based on the nonparametric (local linear) model and 0.37 NCE points based on the parametric (cubic) model, compared with an impact estimate of -0.14 in the baseline nonparametric model and -0.07 in the baseline parametric model. In each case, the impact estimate was not statistically significant. This difference of 0.44 NCE points between the RD impact estimates with manipulation and the baseline RD estimate (for both the parametric and nonparametric specifications) can be translated into an estimate of the RD bias due to manipulation of 0.02 effect size (or standard deviation) units. In this scenario, this amount and type of manipulation had a small effect on the impact estimates—the estimates with manipulation were quite similar to the baseline estimates.

Figure C.1. Density of Pretest Scores in Original Dataset (Ed Tech)



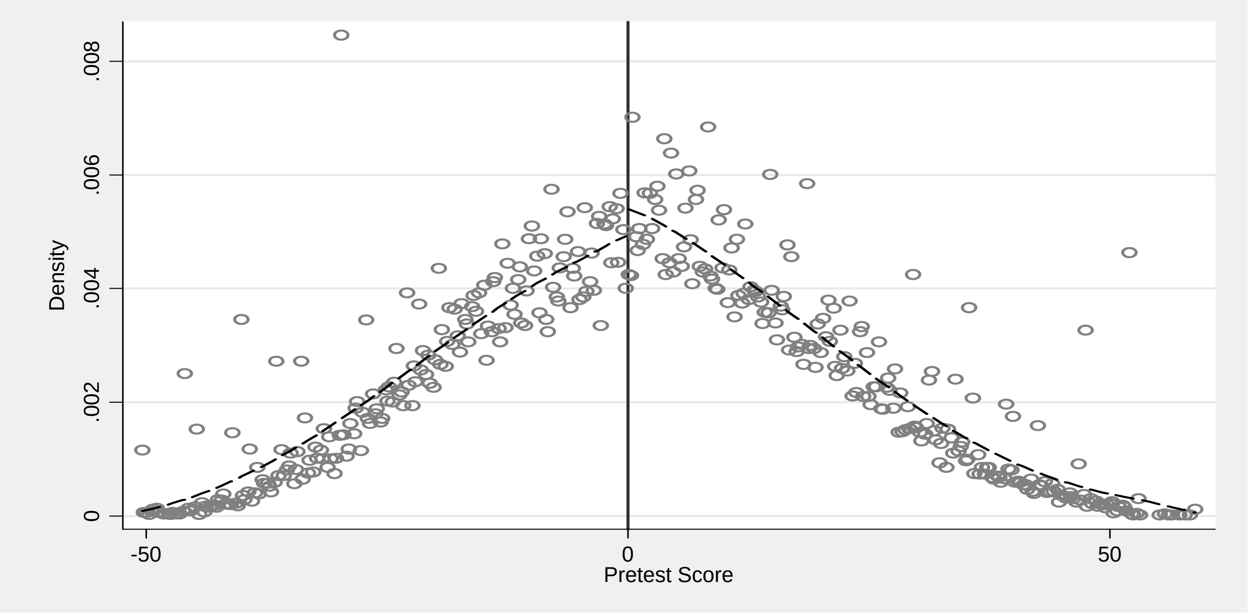
Estimate of discontinuity at cutoff = -0.01

McCrary Test Z-stat = -0.30

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

Figure C.2. Density of Pretest Scores in Manipulated Dataset (Ed Tech) 3% Cheating, 15 Point Window, Cheating is Correlated with Outcome



Estimate of discontinuity at cutoff = 0.09

McCrary Test Z-stat = 1.86*

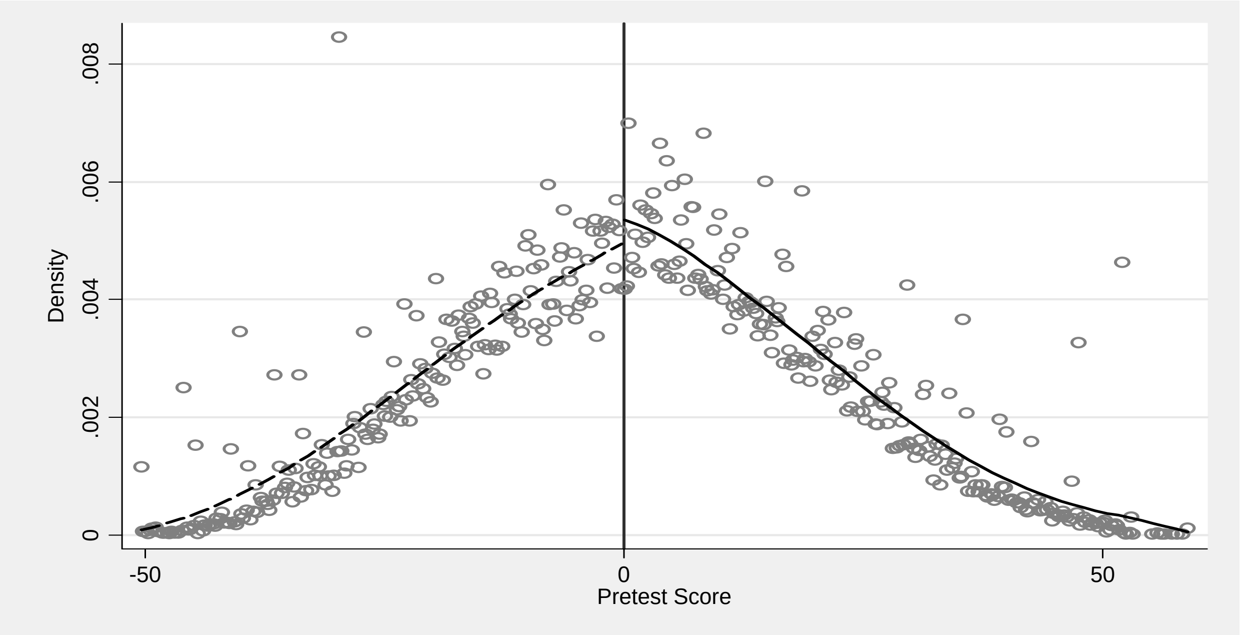
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

Figure C.3 and column 3 of Table C.1 use the same fraction cheating and window, but show the case where cheating was not related to the outcome. This scenario is less likely to lead to bias, because it is no longer true that manipulated observations are more likely to have high posttest scores, conditional on their pretest score. Under this scenario, the McCrary test z-statistic was 1.56, so this data set would also pass the McCrary test. The impact estimates of -0.24 and -0.27 for the local linear and parametric specifications were again similar to one another, as well as being similar to the baseline RD impact estimate. None of the estimates was statistically significant, and the estimated RD bias due to manipulation was small at -0.01 to 0.00 effect size units.

Tables C.2 and C.3 and Figures C.4 through C.9 present the corresponding specifications for the TFA Math and TFA Reading samples. These results are reasonably consistent with the Ed Tech results presented in Table C.1. Table C.2 presents the TFA Math results for the two scenarios with 3 percent cheating and a 15-point manipulation window. For example, the baseline model indicates that the estimated impact of TFA on math scores based on the local linear specification was 3.51 NCE points, while the local linear impact estimates for the two scenarios with manipulation were 4.53 and 3.24 NCE points. All three estimates were statistically significant at the .01 level. The RD bias due to manipulation was 1.02 NCE point or 0.05 effect size units in the scenario in which cheating was correlated with the outcome and -0.17 NCE points or -0.01 effect size units in the no correlation scenario. The parametric estimates also showed little RD bias due to manipulation with estimates of bias in effects size unit of -0.01 to 0.02. The TFA Reading results in Table C.3 showed smaller estimates of RD bias due to manipulation, ranging from -0.01 to 0.01 effect size units. All TFA specifications shown in these two tables passed the McCrary test, so manipulation would not be detected.

Figure C.3. Density of Pretest Scores in Manipulated Dataset (Ed Tech) 3% Cheating, 15 Pt Window, Cheating not Correlated with Outcome



Estimate of discontinuity at cutoff = 0.08

McCrary Test Z-stat = 1.56

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

Table C.2. Regression Results Using Manipulated Data (TFA Math)

	Baseline Model[1]	With Manipulation	
	(1)	(2)	(3)
Fraction Cheating	n/a	3%	3%
Manipulation Window	n/a	15 points	15 points
Is cheating correlated with outcome?	n/a	Yes	No
McCrary Test Z-stat	0.26	0.80	0.79
Local Linear Estimates			
Optimal Bandwidth	6.40	5.83	6.42
Impact Estimate	3.51***	4.53***	3.34***
	(0.85)	(0.90)	(0.85)
Sample Size	3262	2956	3222
Bias due to manipulation (ES)		0.05	-0.01
Parametric Estimates			
Impact Estimate	2.43***	2.87***	2.31***
	(0.64)	(0.64)	(0.64)
Sample Size	9730	9618	9618
Bias due to manipulation (ES)		0.02	-0.01

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports selected coefficients from regression models that also included random assignment block fixed effects and a classroom random effect. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[1] The baseline model is the preferred RD model for the non-parametric or parametric specification using simulated data.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Table C.3. Regression Results Using Manipulated Data (TFA Reading)

	Baseline Model[1]	With Manipulation	
	(1)	(2)	(3)
Fraction Cheating	n/a	3%	3%
Manipulation Window	n/a	15 points	15 points
Is cheating correlated with outcome?	n/a	Yes	No
McCrary Test Z-stat	-0.39	-0.03	0.29
Local Linear Estimates			
Optimal Bandwidth	13.66	12.57	13.72
Impact Estimate	0.91	1.10*	0.78
	(0.59)	(0.62)	(0.59)
Sample Size	5906	5423	5819
Bias due to manipulation (ES)		0.01	-0.01
Parametric Estimates			
Impact Estimate	0.50	0.68	0.40
	(0.68)	(0.68)	(0.68)
Sample Size	9776	9659	9672
Bias due to manipulation (ES)		0.01	0.00

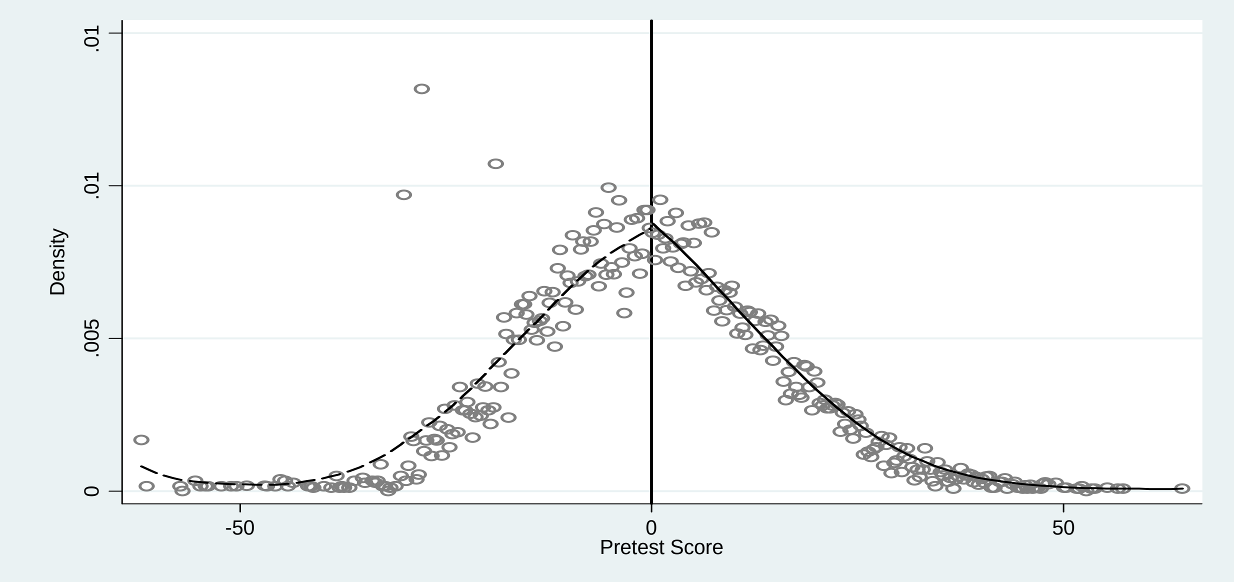
Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports selected coefficients from regression models that also included random assignment block fixed effects and a classroom random effect. Standard errors are shown in parentheses. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

[1] The baseline model is the preferred RD model for the non-parametric or parametric specification using simulated data.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

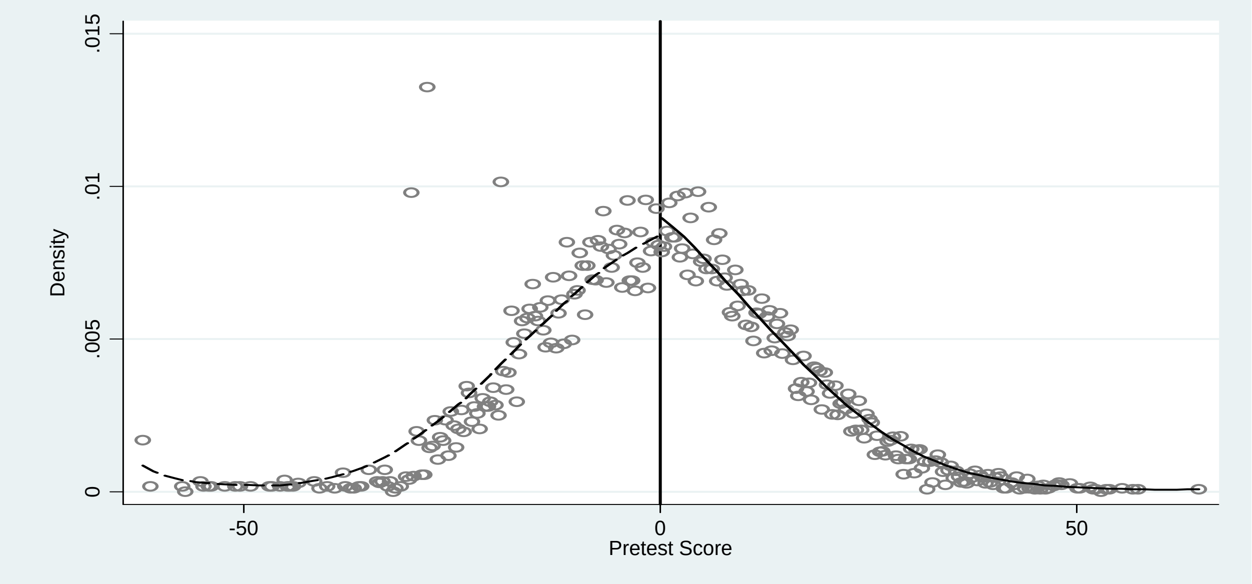
Figure C.4. Density of Pretest Scores in Original Dataset (TFA Math)



Estimate of discontinuity at cutoff = 0.02

McCrary Test Z-stat = 0.26

Figure C.5. Density of Pretest Scores in Manipulated Dataset (TFA Math) 3% Cheating, 15 Point Window, Cheating is Correlated with Outcome



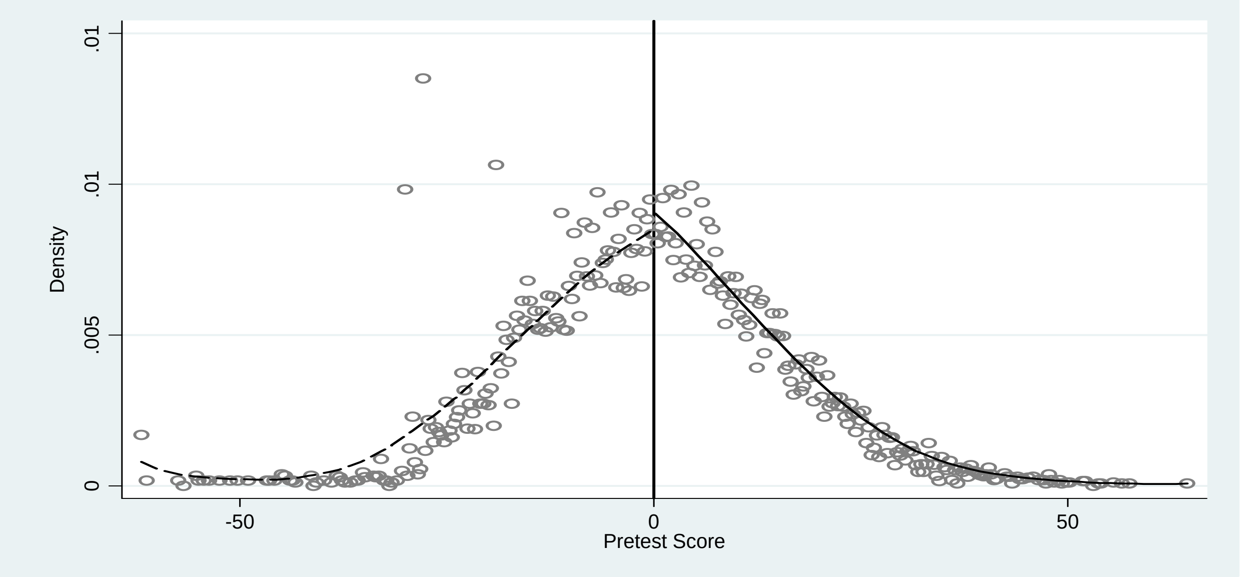
Estimate of discontinuity at cutoff = 0.07

McCrary Test Z-stat = 0.80

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

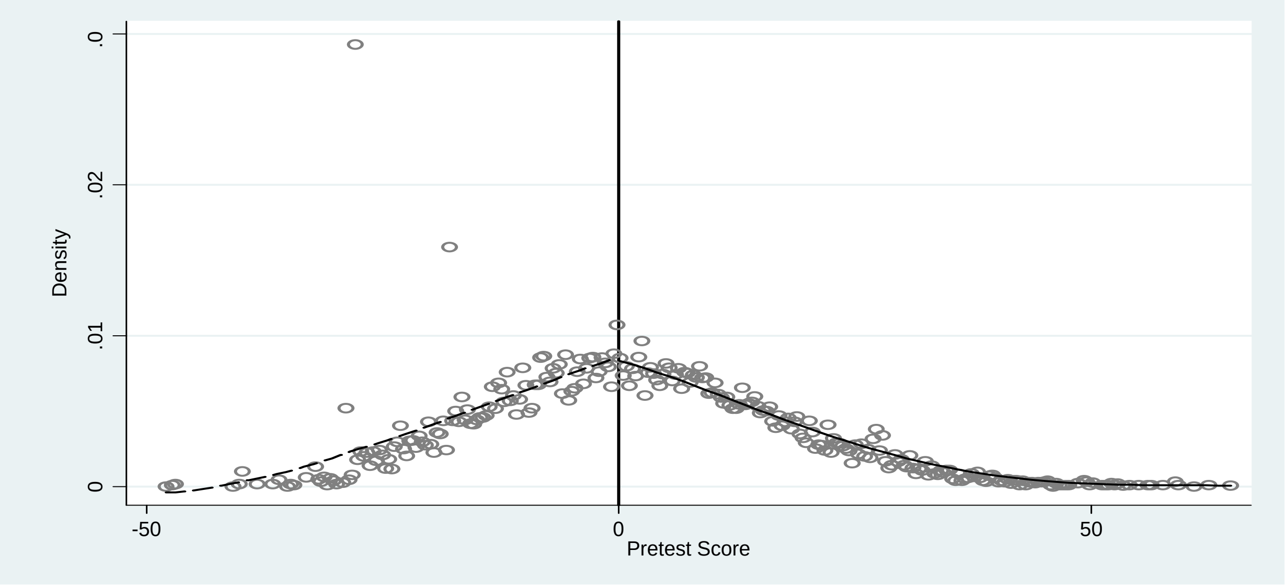
Figure C.6. Density of Pretest Scores in Manipulated Dataset (TFA Math) 3% Cheating, 15 Pt Window, Cheating not Correlated with Outcome



Estimate of discontinuity at cutoff = 0.07

McCrary Test Z-stat = 0.79

Figure C.7. Density of Pretest Scores in Original Dataset (TFA Read)



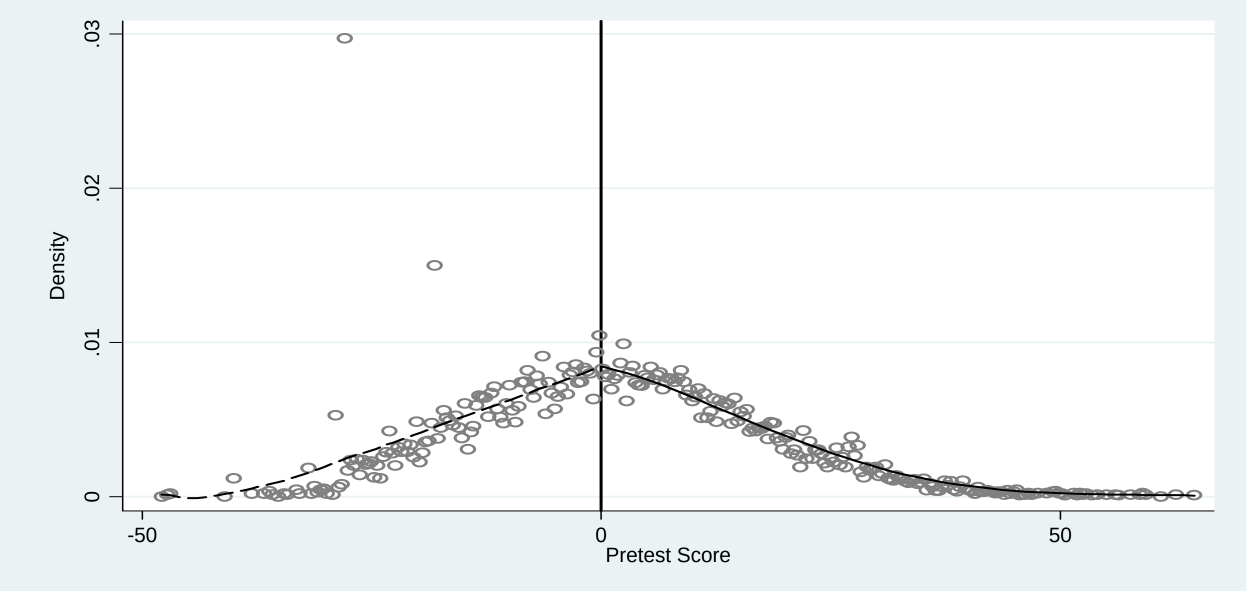
Estimate of discontinuity at cutoff = -0.03

McCrary Test Z-stat = -0.39

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

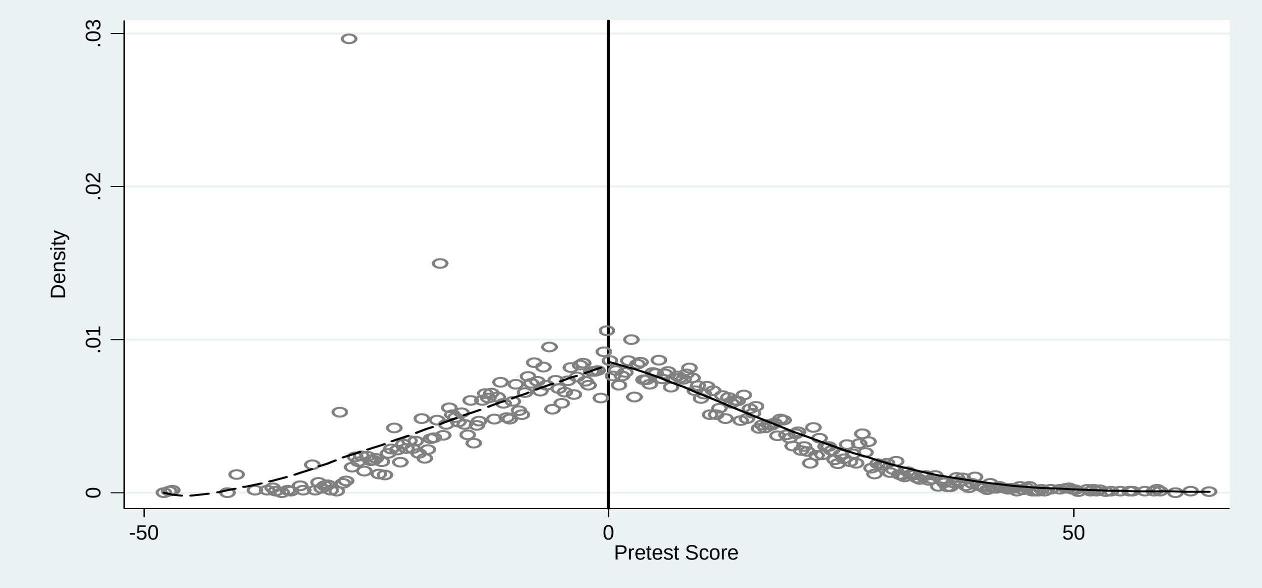
Figure C.8. Density of Pretest Scores in Manipulated Dataset (TFA Read) 3% Cheating, 15 Point Window, Cheating is Correlated with Outcome



Estimate of discontinuity at cutoff = -0.00

McCrary Test Z-stat = -0.03

Figure C.9. Density of Pretest Scores in Manipulated Dataset (TFA Read) 3% Cheating, 15 Pt Window, Cheating not Correlated with Outcome



Estimate of discontinuity at cutoff = 0.02

McCrary Test Z-stat = 0.29

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Each circle represents the density (shown on the y axis) of observations falling within that cell of the assignment variable (shown on the x axis). The dashed lines are local polynomial approximations of the density on each side of the cutoff.

The remaining tables and figures summarize results for the 200 different manipulation scenarios we examined for each original data set. The next several paragraphs describe the contents of the tables and figures; we then discuss the patterns of results.

Tables C.4 through C.6 summarize the results for different values of the fraction of control students cheating, holding the manipulation window constant at 15 points, for the Ed Tech (Table C.4), TFA Math (Table C.5) and TFA Reading (Table C.6) samples. Each table shows the McCrary test statistic and estimates of RD bias due to manipulation for the local linear and parametric specifications, with the top panel presenting these estimates for the case where cheating is correlated with the outcome and the bottom panel showing the case where there is no correlation. Note that the results shown in row 3 of each panel represent the same scenarios shown in Tables C.1 through C.3. Moving down from row 1 through row 10 of the top panel for each dataset—from smaller to larger fractions cheating, both the McCrary test statistic and the size of the bias increase.

Tables C.7 through C.9 summarize models where the fraction cheating was held constant at 3 percent, but the window from which manipulated observations were drawn varies from 5 to 50 points. In these tables, the scenarios in which RD bias due to manipulation was largest were those in which manipulation was correlated with the outcome and the manipulation window was small - 10 points or less. For the Ed Tech data, the McCrary test also indicated a significant amount of manipulation. In these scenarios with a small manipulation window, students with posttest scores above the median for their pretest score were being moved from right below the cutoff pretest score to right above the cutoff. This movement led to substantial RD bias as well as a significant discontinuity in the density of pretest scores detected by the McCrary test.

Figures C.10 through C.21 present information from these same 200 manipulation scenarios graphically, showing the RD bias due to manipulation for each pair of values indicating the fraction cheating and size of the manipulation window. Based on the classification of the magnitudes of effect sizes of education interventions in the literature discussed in chapter 5, we defined three bias categories. We defined low RD bias due to manipulation as a difference of less than 0.05 effect size units, medium bias as 0.05 to 0.15 effect size units, and high bias as a difference greater than 0.15 effect size units. For each manipulation scenario summarized in these figures, a different symbol indicates whether the RD bias due to manipulation was low, medium, or high.

These figures also show, for each scenario, whether the results of the McCrary test indicated the presence of manipulation. In particular, the shaded area covers scenarios where the manipulated data set passed the McCrary test (that is, the test failed to detect the presence of manipulation). In the non-shaded areas of these figures, the manipulated data sets failed the McCrary test (that is, the test did detect the presence of manipulation).

There are two figures for each scenario/specification. The first figure shows the case within which cheating was correlated with the outcome; the second shows the case without this correlation. The patterns present in Tables C.4 through C.9 are also evident in these figures. Both the local linear and parametric specifications performed well, except in cases where there was a high fraction cheating and a small manipulation window.

Looking at the full set of manipulation scenarios for the three data sets - Ed Tech, TFA Math and TFA Reading – four main patterns of results emerged and are illustrated in Tables C.4 through C.9 and Figures C.10 through C.21.

When the McCrary test failed to show evidence of manipulation of the assignment variable, RD bias from manipulation was usually low (less than 0.05 effect size units). This can be seen by looking at the bias markers in the shaded area of each figure, which represent manipulation scenarios that pass the McCrary test. In most of these scenarios the bias was less than 0.05 effect size units. The exceptions were in the TFA data sets in cases where there was a large fraction cheating and a small manipulation window. In manipulation scenarios in which the McCrary test showed evidence of manipulation, shown in the non-shaded areas of these graphs, the RD bias due to manipulation was low in some cases and high in others.

When manipulation of the assignment variable was not correlated with the posttest outcome measure (conditional on the pretest score), RD bias due to manipulation was low. An example of this pattern can be seen in the odd-numbered figures. Nearly all manipulation scenarios produced low bias in these cases. The exceptions included one manipulation scenario for the Ed Tech local linear specification and three scenarios for the TFA Math local linear specifications. In each of these four scenarios, eight or more percent of the sample cheated and the manipulation window was 10 points or less. This would be a quite extreme case of manipulation. In the Ed Tech case (with a 5 point window), for example, this would amount to manipulation of the pretest scores of about 27 percent of the students in the relevant window to allow them to receive the treatment intervention.

When manipulation of the assignment variable was correlated with the posttest outcome measure (conditional on the pretest score), RD bias due to manipulation was dependent on (a) the amount of cheating among control students; and (b) the size of the window from which the manipulation arose. Substantial bias was most likely when there was both a large amount of cheating and the manipulated observations came from cases close to the pretest cutoff score. In cases when cheaters came from nearly the entire pretest distribution, bias tended to be low. When 3 percent (or less) of control cases cheated, bias also tended to be low, unless manipulation was restricted to observations immediately below the cutoff.[59]

RD bias due to manipulation was not strongly affected by the specification of the RD model that was estimated (that is, whether it was a parametric or nonparametric model). This pattern can be seen by comparing, for example, Figures C.10 and C.12 for the Ed Tech data set. These two graphs present the manipulation scenarios where cheating is correlated with the outcome for the local linear (Figure C.10) and parametric (Figure C.12) specifications. The pattern of bias was nearly identical for these two specifications.

The McCrary test results (and the corresponding shading) were identical across the two figures, which will always be true since the McCrary test only uses pretest scores. There were some notable differences between the Ed Tech and TFA results. First, the McCrary test nearly always failed to detect manipulation in the case of the TFA data (in all but 74 out of 800 scenarios), while more frequently detecting manipulation in the case of the Ed Tech data. Second, there were a number of scenarios with TFA data sets that passed the McCrary test but that produced impact estimates in which the RD bias due to manipulation was moderate rather than low. In contrast, all Ed Tech data sets that passed the McCrary test had low bias.

We believe that these differences are due to the fact that the McCrary test had lower statistical power when based on the TFA data, with its smaller sample sizes. Thus, the evidence of a discontinuity in the density of the pretest score at the cutoff would have to be larger in order for the test to detect manipulation in the case of the TFA data than in the case of the Ed Tech data. By contrast, the classification of RD bias due to manipulation into the low, medium, and high categories was based on a non-statistical procedure that did not account for sampling variability.

The results of our manipulation exercise suggest that a small degree of manipulation of the assignment variable in an RD study is unlikely to bias RD impact estimates substantially, and that if the McCrary test fails to detect manipulation (and has sufficient statistical power), the RD impact estimates are likely to show low bias. The results generally matched our expectations of how bias would vary with the manipulation parameters. For both the local linear and parametric specifications, bias increases with the fraction cheating, is larger when the window from which manipulation occurs is small, and is larger when the likelihood of cheating is correlated with an individual's posttest score. Both the local linear and parametric specifications showed low bias in most variations of manipulation that we tested. The problematic cases occurred when pretest scores were manipulated for a large fraction of students who are likely to benefit from the treatment.

[59] 3% cheating represents a sizeable fraction of the available observations in the 5-point manipulation window. For Ed Tech it is about 10% of the eligible observations.

Table C.4. Bias Relative to Baseline Model for Varying Fraction Cheating (Ed Tech) Fixed Parameters: 15 Point Manipulation Window

Cheating Is Correlated With Outcome		Bias Due To Manipulation	
Fraction Cheating	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
1	.034	0.01	0.00
2	1.14	0.02	0.01
3	1.86*	0.02	0.02
4	2.43**	0.03	0.02
5	3.07***	0.04	0.04
6	3.90***	0.06	0.05
7	4.50***	0.05	0.05
8	4.99***	0.06	0.05
9	5.39***	0.07	0.06
10	6.21***	0.11	0.09

Cheating Is Not Correlated With Outcome			
Fraction Cheating	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
1	0.37	0.00	0.00
2	1.11	0.00	0.00
3	1.56	0.00	-0.01
4	2.28**	0.00	-0.01
5	3.21***	0.00	-0.01
6	4.00***	0.01	-0.01
7	4.47***	0.00	-0.02
8	4.98***	0.01	-0.02
9	5.32***	0.01	-0.01
10	6.28***	0.02	0.00

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Bias due to manipulation is reported in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Table C.5. Bias Relative to Baseline Model for Varying Fraction Cheating (TFA Math) Fixed Parameters: 15 Point Manipulation Window

Cheating Is Correlated With Outcome		Bias Due To Manipulation	
Fraction Cheating	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
1	0.59	0.04	0.01
2	0.67	0.03	0.02
3	0.80	0.05	0.02
4	0.90	0.04	0.04
5	1.17	0.03	0.04
6	1.22	0.06	0.06
7	1.49	0.07	0.06
8	1.51	0.10	0.07
9	1.56	0.08	0.09
10	2.15**	0.10	0.10
Cheating Is Not Correlated With Outcome			
Fraction Cheating	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
1	0.47	0.00	0.00
2	0.61	-0.01	0.00
3	0.79	-0.01	-0.01
4	1.02	0.00	0.00
5	1.17	0.01	0.01
6	1.43	0.01	0.02
7	1.33	-0.02	-0.02
8	1.65*	0.03	0.00
9	1.80*	0.00	-0.01
10	1.92*	0.02	0.01

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Bias due to manipulation is reported in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Table C.6. Bias Relative to Baseline Model for Varying Fraction Cheating (TFA Reading) Fixed Parameters: 15 Point Manipulation Window

Cheating Is Correlated With Outcome		Bias Due To Manipulation	
Fraction Cheating	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
1	-0.01	0.00	0.00
2	0.07	0.01	0.01
3	-0.03	0.01	0.01
4	0.61	0.03	0.02
5	0.50	0.03	0.02
6	0.67	0.03	0.03
7	1.33	0.04	0.04
8	0.97	0.05	0.05
9	1.38	0.05	0.06
10	1.50	0.05	0.07
Cheating Is Not Correlated With Outcome			
Fraction Cheating	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
1	0.03	-0.01	0.00
2	0.09	0.01	0.01
3	0.29	-0.01	0.00
4	0.60	-0.01	-0.01
5	0.53	0.00	0.01
6	0.72	-0.01	0.00
7	0.85	-0.01	0.00
8	0.78	0.00	0.00
9	1.42	0.01	0.02
10	0.90	-0.01	-0.01

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Bias due to manipulation is reported in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Table C.7. Bias Relative to Baseline Model for Varying Manipulation Window (Ed Tech) Fixed Parameters: 3% of Sample Cheating

Cheating Is Correlated With Outcome		Bias Due To Manipulation	
Manipulation Window (points)	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
5	3.45***	0.11	0.09
10	2.52**	0.05	0.05
15	1.86*	0.02	0.02
20	1.46	0.02	0.00
25	1.36	0.01	-0.01
30	1.55	0.00	-0.02
35	1.38	0.00	-0.02
40	1.41	0.00	-0.02
45	1.24	0.00	-0.02
50	1.46	-0.01	-0.02
Cheating Is Not Correlated With Outcome			
Manipulation Window (points)	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
5	3.21***	0.02	0.02
10	2.35**	0.01	0.00
15	1.56	0.00	-0.01
20	1.51	0.00	-0.02
25	1.39	-0.01	-0.02
30	1.48	-0.01	-0.03
35	1.24	-0.02	-0.04
40	1.14	-0.02	-0.04
45	1.32	-0.01	-0.03
50	1.36	-0.01	-0.03

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Bias due to manipulation is reported in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Table C.8. Bias Relative to Baseline Model for Varying Manipulation Window (TFA Math) Fixed Parameters: 3% of Sample Cheating

Cheating Is Correlated With Outcome		Bias Due To Manipulation	
Manipulation Window (points)	McCrary Test z-stat	Local Linear	Parametric (Cubic)
5	1.65*	0.12	0.07
10	1.01	0.03	0.04
15	0.80	0.05	0.02
20	0.78	0.04	0.02
25	0.73	0.04	0.02
30	0.60	0.02	0.01
35	0.82	0.04	0.01
40	0.89	0.00	-0.01
45	0.59	0.03	0.02
50	0.59	0.02	0.01
Cheating Is Not Correlated With Outcome			
Manipulation Window (points)	McCrary Test z-stat	Local Linear	Parametric (Cubic)
5	1.41	-0.01	0.00
10	0.98	0.02	0.01
15	0.79	-0.01	-0.01
20	0.93	-0.01	-0.01
25	0.71	0.00	0.00
30	0.65	0.00	0.00
35	0.63	0.00	0.00
40	0.90	0.00	-0.02
45	0.87	0.03	-0.01
50	1.02	-0.02	-0.02

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Bias due to manipulation is reported in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Table C.9. Bias Relative to Baseline Model for Varying Manipulation Window (TFA Reading) Fixed Parameters: 3% of Sample Cheating

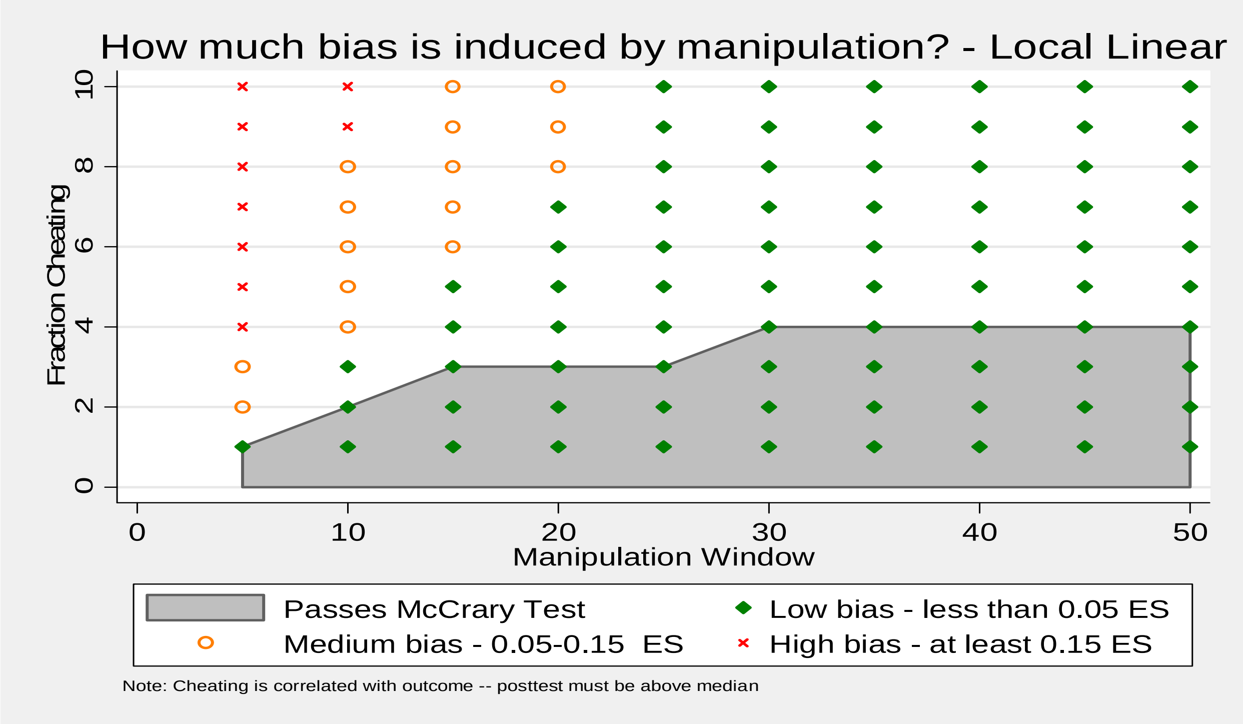
Cheating is Correlated with Outcome		Bias Due to Manipulation	
Manipulation Window (points)	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
5	0.77	0.07	0.07
10	0.74	0.04	0.03
15	-0.03	0.01	0.01
20	0.14	0.02	0.02
25	0.28	0.01	0.01
30	0.10	0.01	0.02
35	0.07	-0.01	0.00
40	0.09	0.01	0.01
45	0.01	0.00	0.00
50	0.15	-0.01	-0.01
Cheating is not Correlated with Outcome			
Manipulation Window (points)	McCrary Test Z-Stat	Local Linear	Parametric (Cubic)
5	0.89	0.00	0.01
10	0.44	0.00	0.00
15	0.29	-0.01	0.00
20	0.09	0.00	0.00
25	0.14	0.00	0.00
30	0.00	0.00	0.00
35	-0.01	-0.02	-0.01
40	-0.08	-0.01	-0.01
45	0.05	0.00	0.00
50	0.54	-0.03	-0.03

Source: Data from the Teach for America Study (Decker et al. 2004).

Note: Bias due to manipulation is reported in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

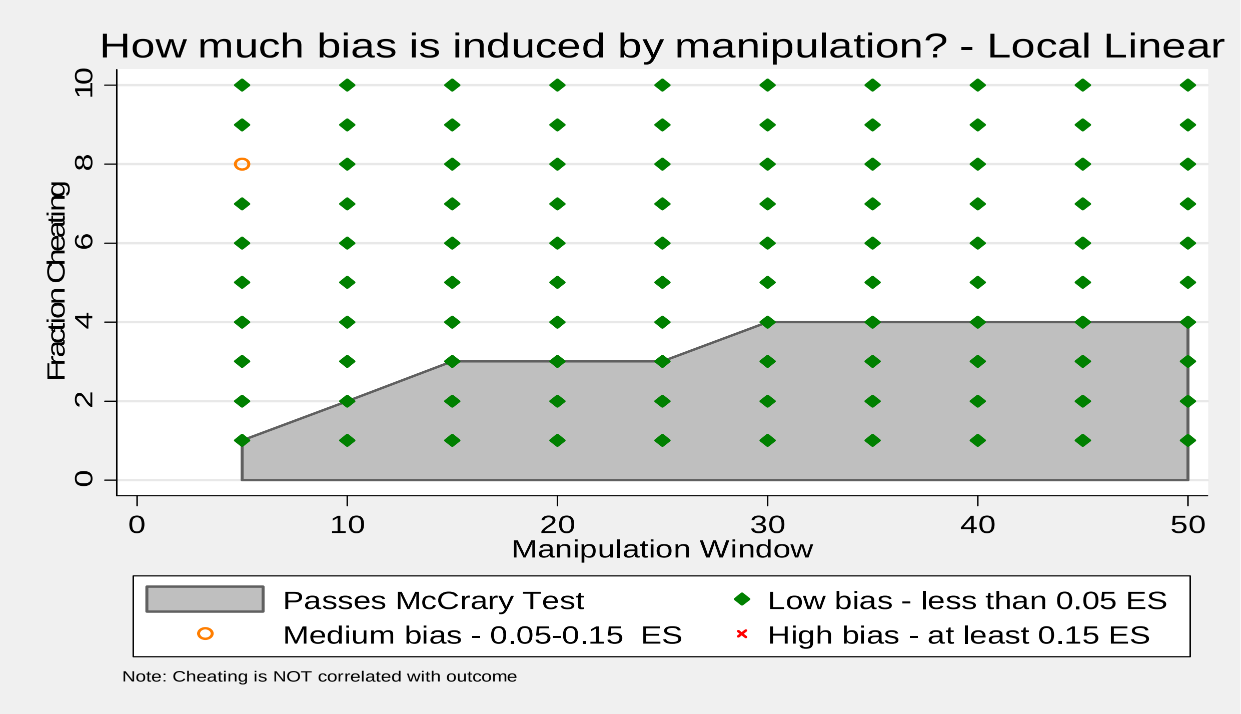
*Significantly different from zero at the .10 level, two-tailed test.
**Significantly different from zero at the .05 level, two-tailed test.
***Significantly different from zero at the .01 level, two-tailed test.

Figure C.10. Summary of Bias for Manipulated Datasets (Ed Tech) Local Linear Specification, Cheating is Correlated with Outcome



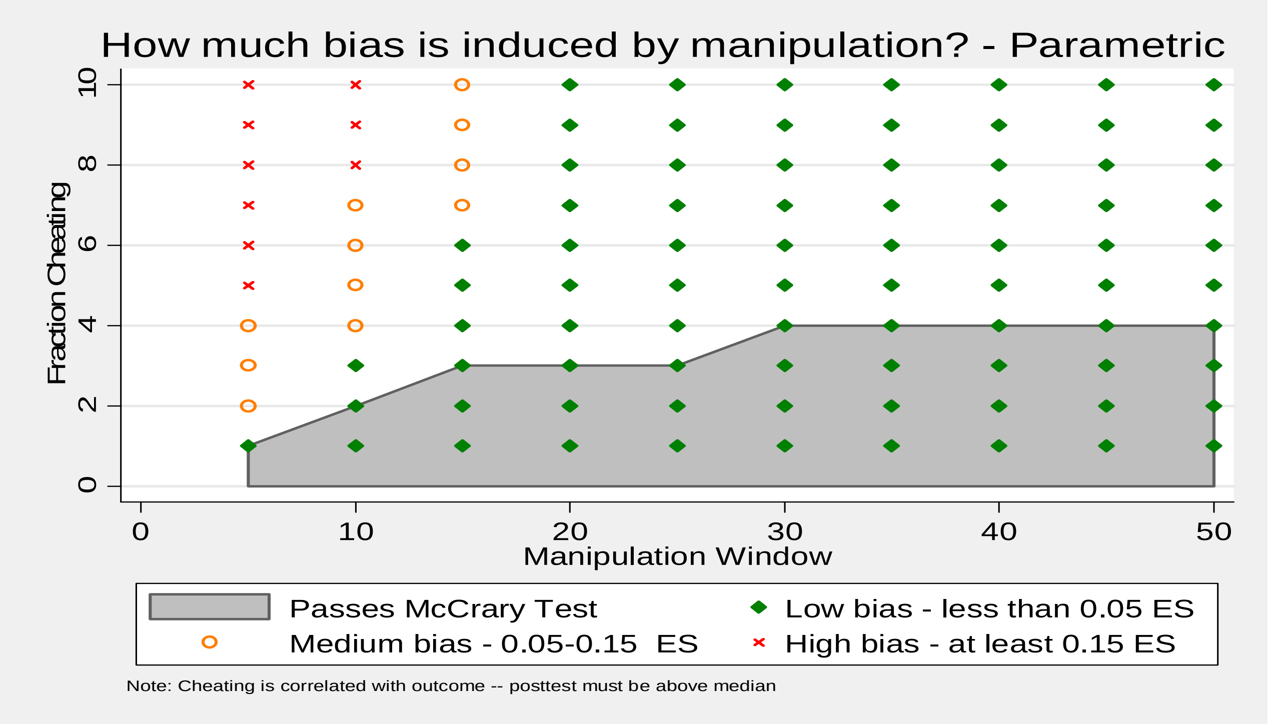
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure C.11. Summary of Bias for Manipulated Datasets (Ed Tech) Local Linear Specification, Cheating Not Correlated with Outcome



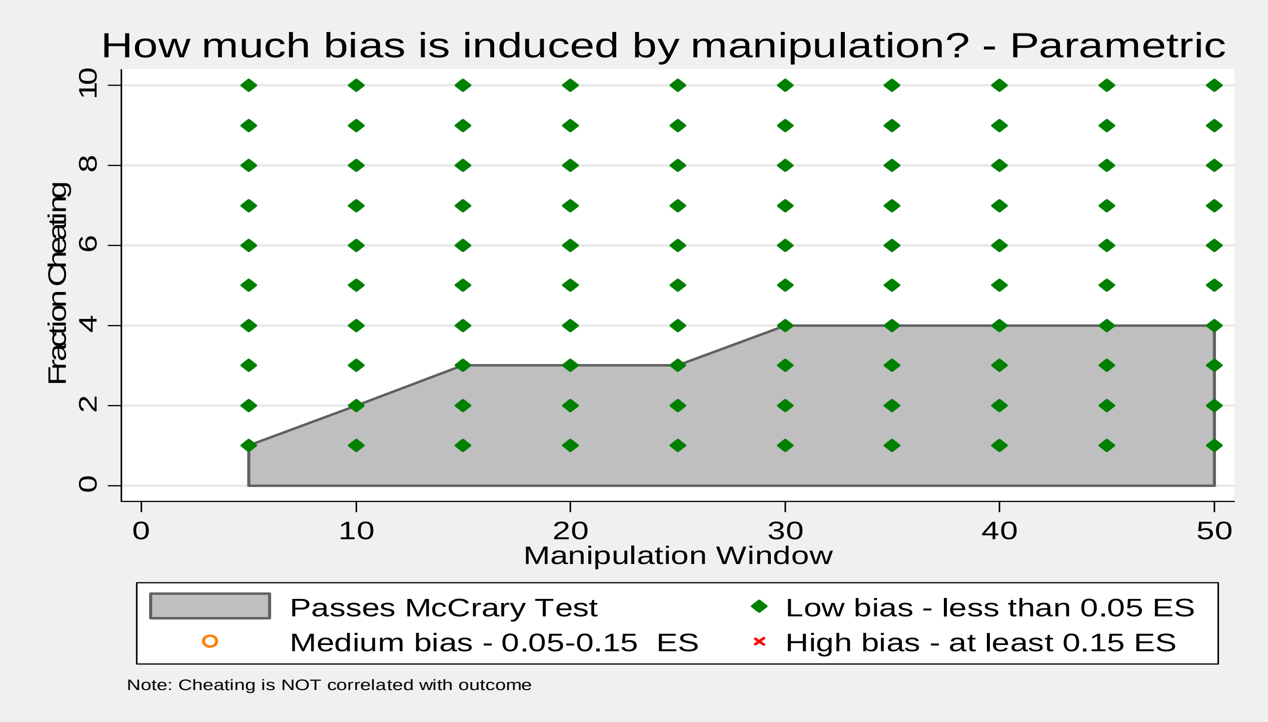
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure C.12. Summary of Bias for Manipulated Datasets (Ed Tech) Parametric Specification, Cheating Is Correlated With Outcome



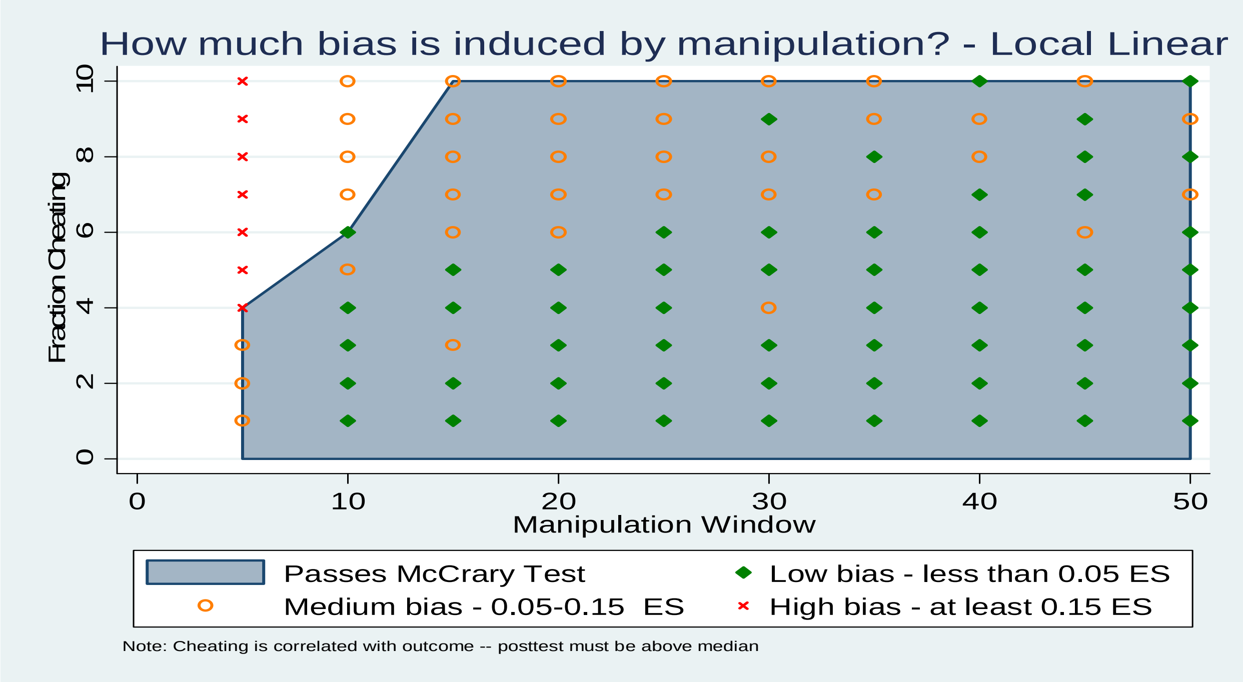
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure C.13. Summary of Bias for Manipulated Datasets (Ed Tech) Parametric Specification, Cheating Not Correlated with Outcome



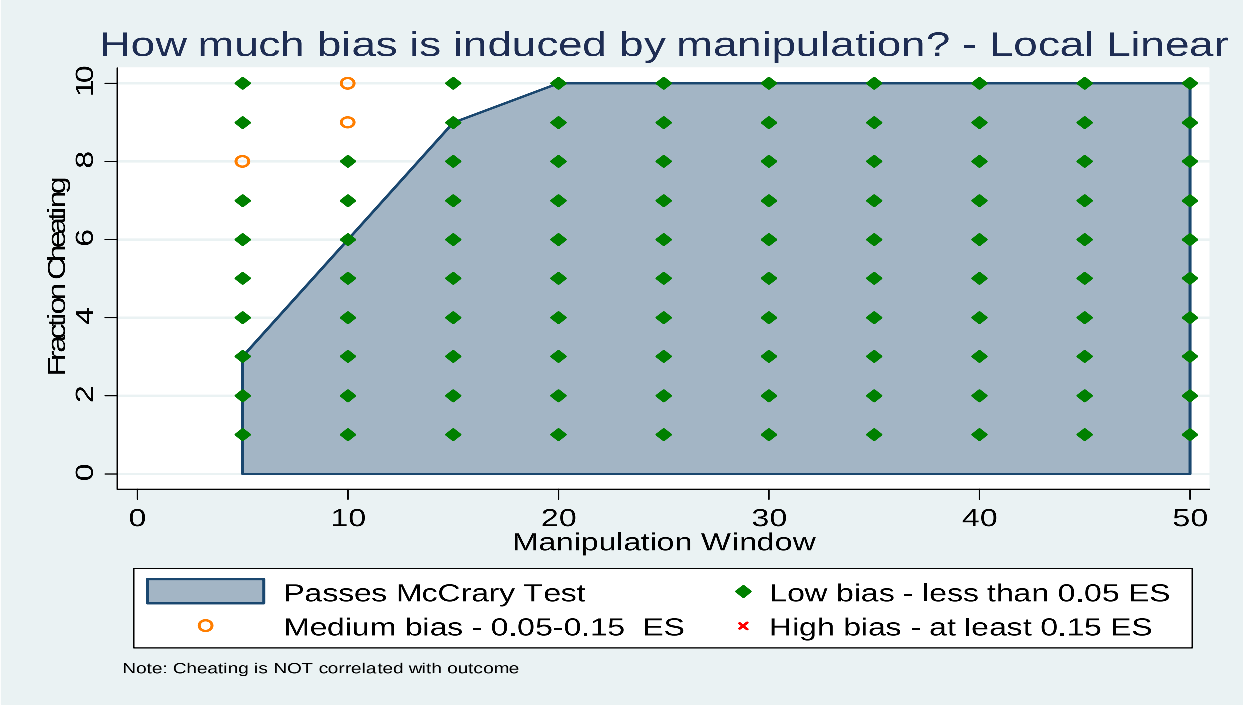
Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Figure C.14. Summary of Bias for Manipulated Datasets (TFA Math) Local Linear Specification, Cheating is Correlated with Outcome



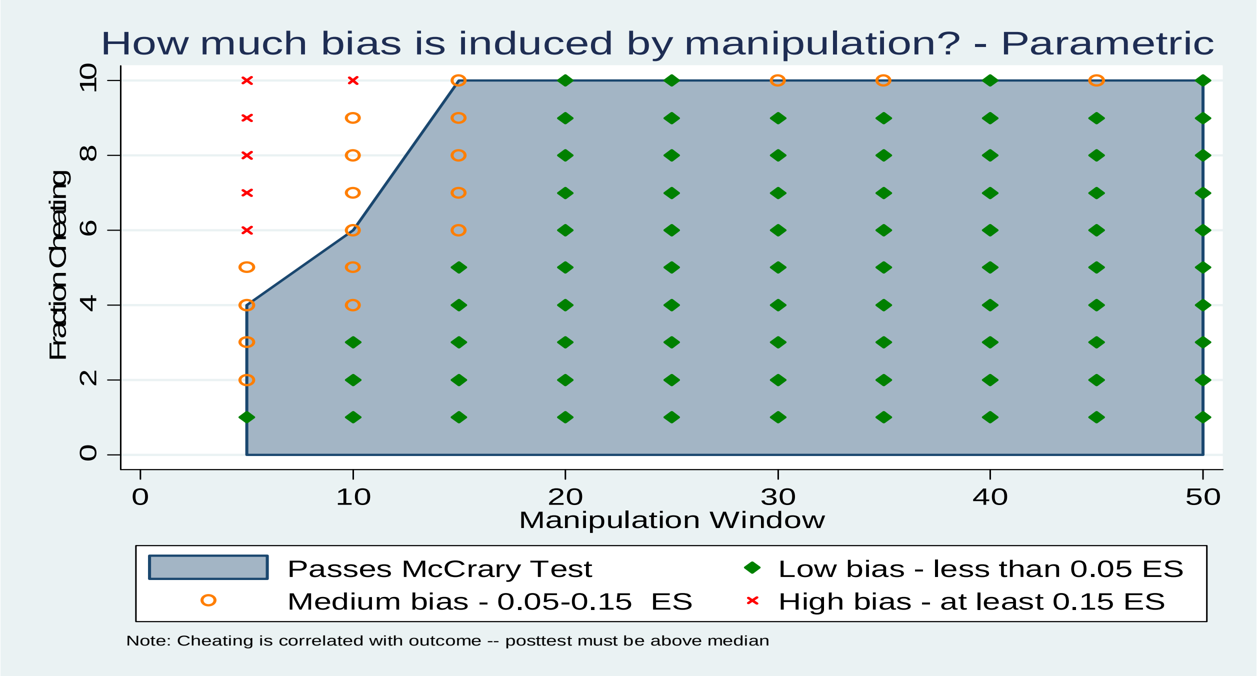
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.15. Summary of Bias for Manipulated Datasets (TFA Math) Local Linear Specification, Cheating Not Correlated with Outcome



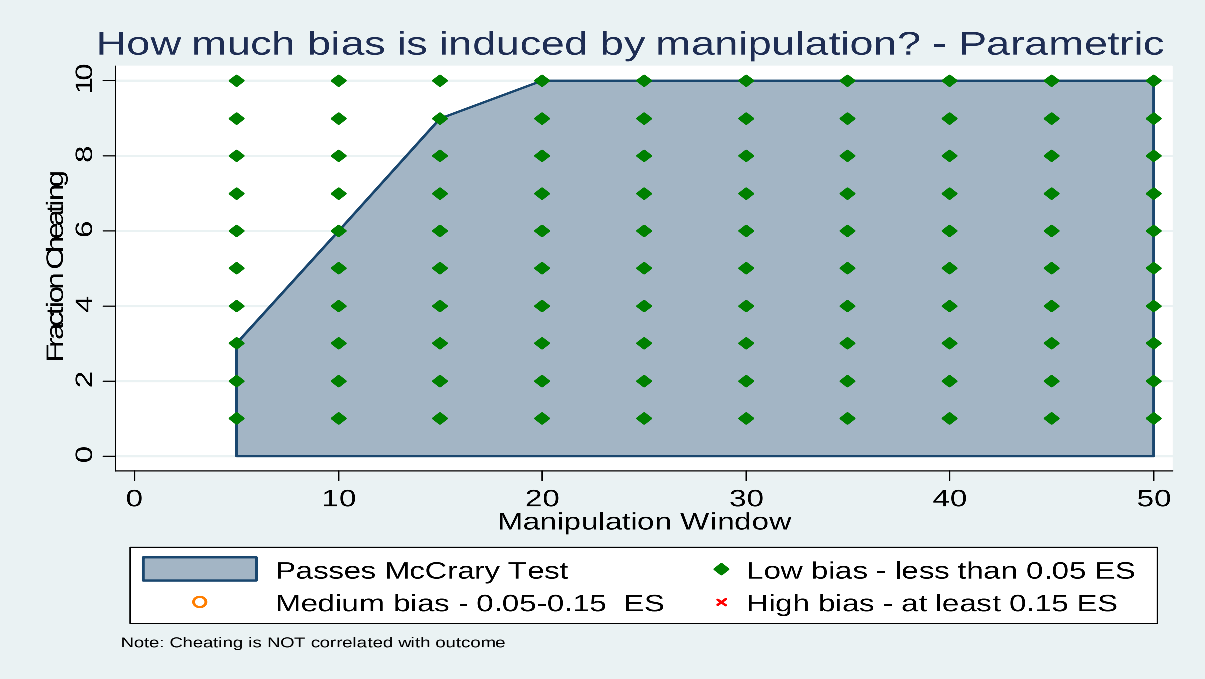
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.16. Summary of Bias for Manipulated Datasets (TFA Math)Parametric Specification, Cheating is Correlated With Outcome



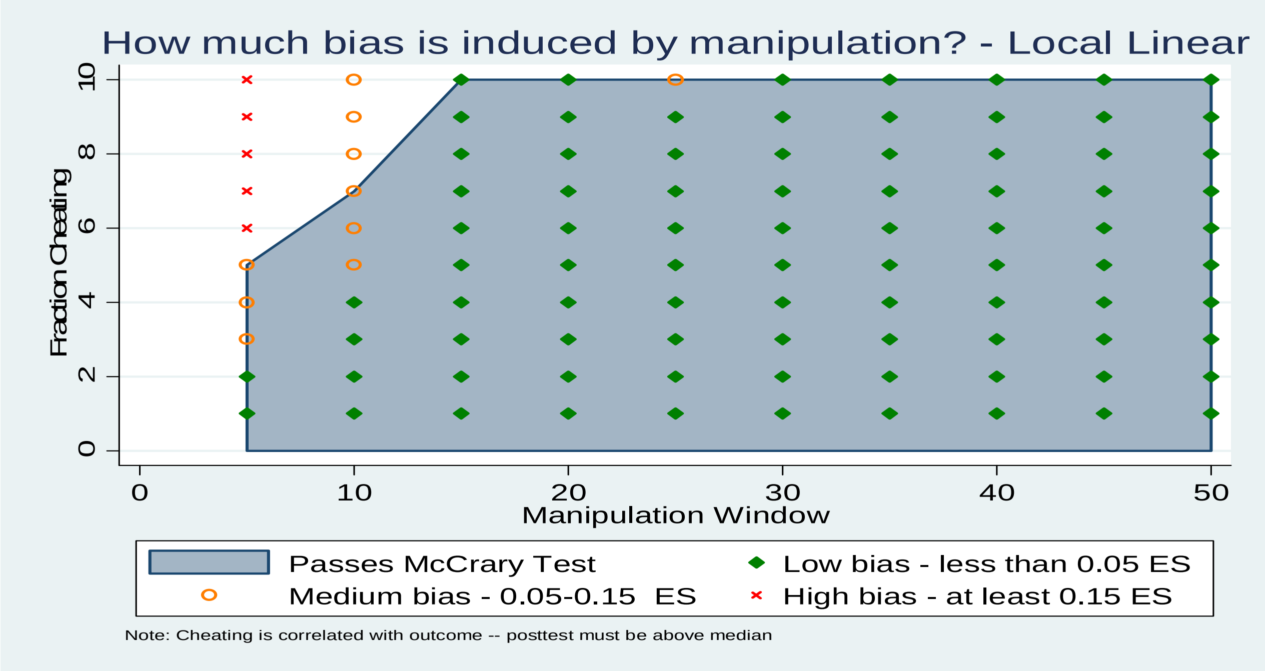
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.17. Summary of Bias for Manipulated Datasets (TFA Math) Parametric Specification, Cheating Not Correlated with Outcome



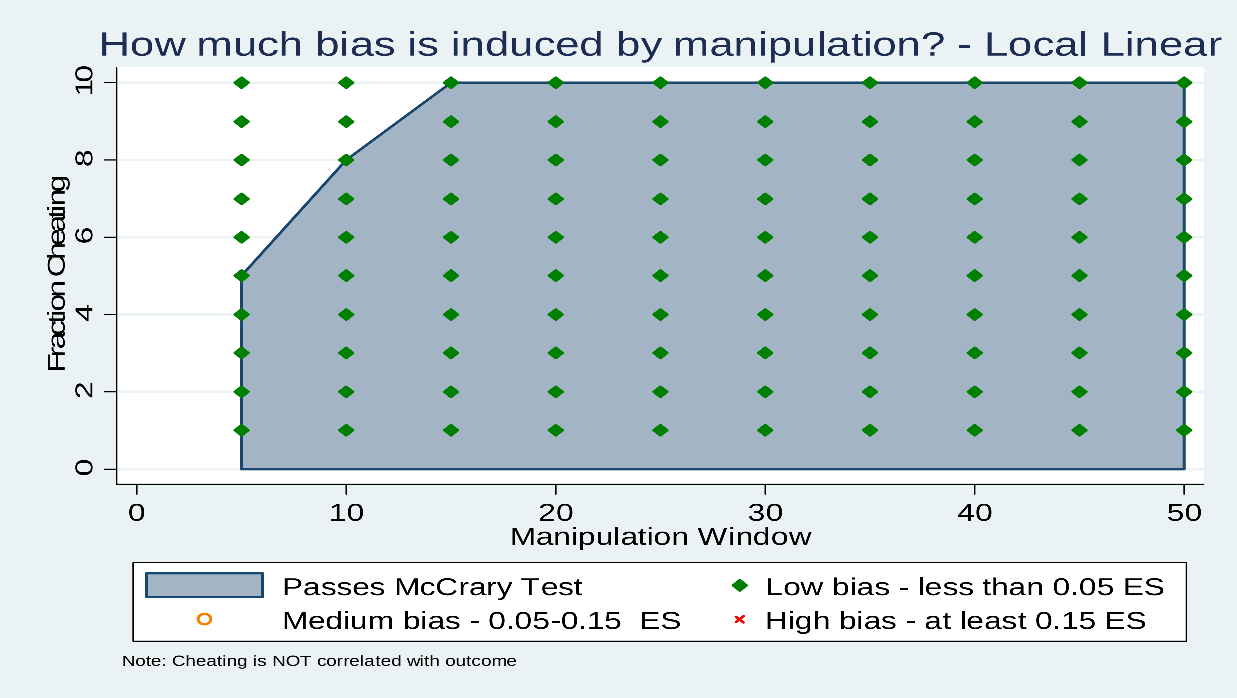
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.18. Summary of Bias for Manipulated Datasets (TFA Read) Local Linear Specification, Cheating is Correlated with Outcome



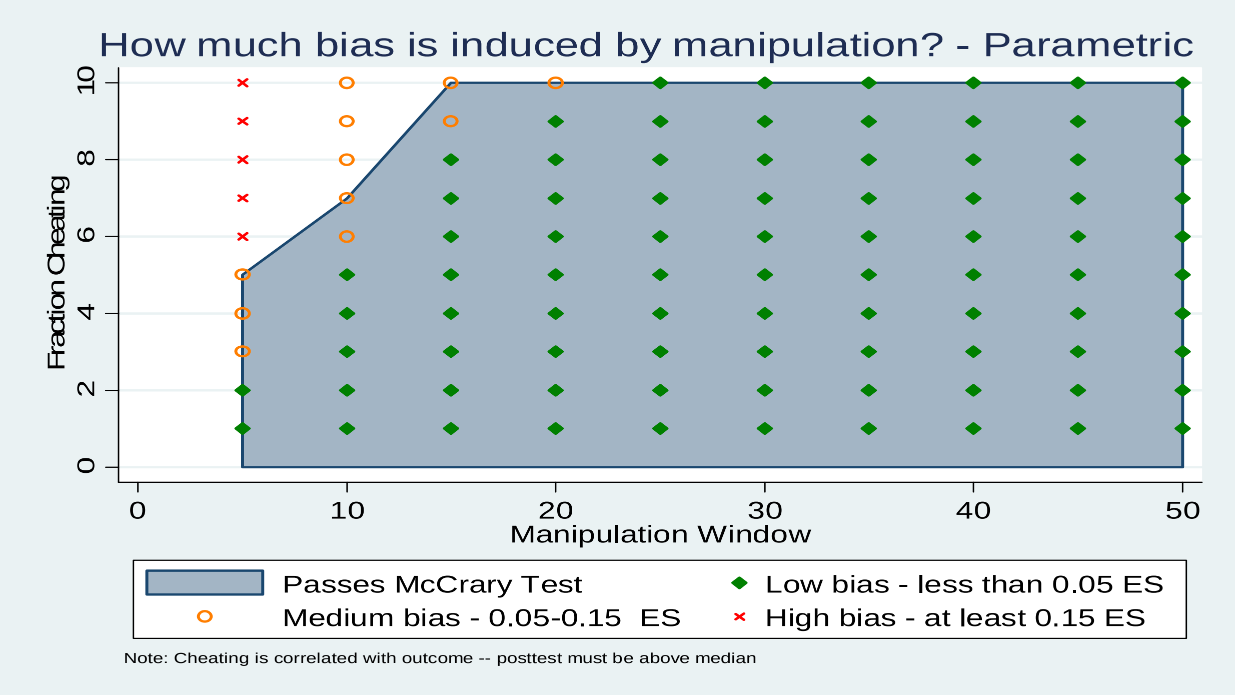
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.19. Summary of Bias for Manipulated Datasets (TFA Read)Local Linear Specification, Cheating Not Correlated with Outcome



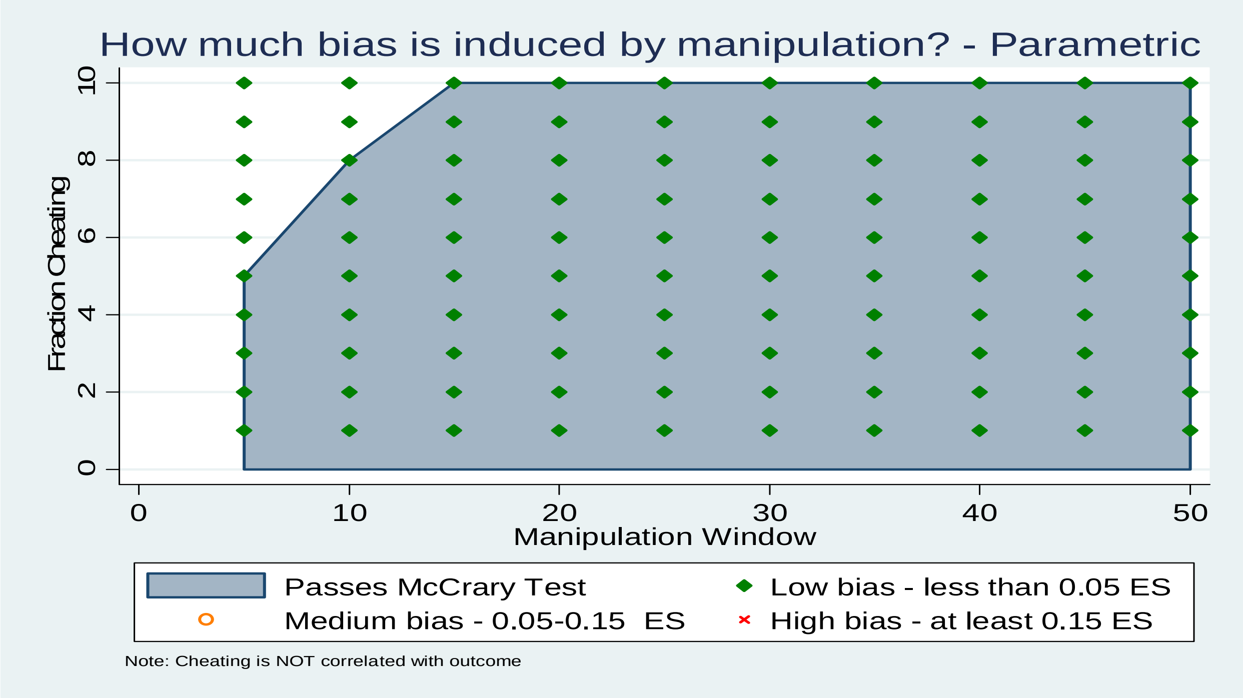
Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.20. Summary of Bias for Manipulated Datasets (TFA Read) Parametric Specification, Cheating is Correlated with Outcome



Source: Data from the Teach for America Study (Decker et al. 2004).

Figure C.21. Summary of Bias for Manipulated Datasets (TFA Read) Parametric Specification, Cheating Not Correlated With Outcome



Source: Data from the Teach for America Study (Decker et al. 2004).

APPENDIX D. SIMULATING DATA FOR THE RD REPLICATION EXERCISE

This appendix describes the steps for simulating additional data to increase the power of the RD analysis and, ultimately, the RD replication exercise. The usual problem with using simulated data to test hypotheses is that in order to simulate new data, one must often make assumptions about the very parameters being estimated. Thus, unless the simulation is carefully conducted, the researcher may assume the answer to the research question being investigated. In the case of simulating data for the RD replication of experimental estimates, our approach to this potential problem involves two components. First, we simulated data using parametric assumptions based on the experimental model and original experimental data set, and then used the simulated data to generate RD High samples and estimate the impacts of Ed Tech, TFA on math scores, and TFA on reading scores using an RD model. Second, in estimating the experimental model used as a basis for simulating the data ultimately used to create the RD High samples, we used a highly flexible set of functional form assumptions; more flexible than the functional form assumptions subsequently used in the RD model. Thus, the simulated values did not artificially reflect a simple assumed parametric relationship between the pretest and posttest values that would then be reflected back in the final RD model estimates.[60] With a flexible functional form, the simulated values more accurately reflected the actual relationship observed in the original experimental data.

The objective of the simulation was to equalize the statistical power of the original experimental model and the RD model. The RD specification had limited power for three reasons: (1) the creation of the RD sample required dropping half of the data; (2) RD models have less statistical power than experimental models because of a correlation between the assignment variable and the treatment status[61]; and (3) the process for choosing an optimal RD specification added variability to the impact estimates. Equalizing the statistical power of the two approaches required significant data generation. To counterbalance these influences, we needed to simulate a sample roughly ten times the size of the original samples.

A. Simulation Process

The process for simulating additional data for the RD replication model included three steps. These are described below, and where relevant, we note differences between the simulation process for the Ed Tech data and the TFA data.

Step One: Estimate the Experimental Model.[62] The first step in the simulation process was to estimate a data-generating model using the full experimental sample. This model specified a

[60] For example, if the simulation model had assumed a linear relationship between the pretest and posttest, the relationship between these two variables in the simulated data would have been linear, implying that the parametric RD estimation process would likely have favored a linear specification and the bandwidth for the local linear RD model would have been wide.

[61] The loss of power in the RD models due to this factor is likely to be even greater in model specifications that include polynomial terms to capture the nonlinear relationship between the assignment variable and outcome.

[62] The experimental model whose results served as the gold standard against which the RD estimates were compared was estimated separately from the experimental model used to generate the additional data for the RD estimates. The gold standard model had a more simple specification, with no additional covariates beyond the pretest score and teacher and block effects.

flexible relationship between the pretest and the posttest allowing for interactions between the pretest and treatment status. We also included an additional covariate that explained a portion of the variation in the outcome. Hypothetically, we could have included additional covariates, but the cost of additional covariates is that they would have complicated the process of simulating the values of the model's full set of covariates.

If *i* is the subscript indexing students, *j* is the subscript indexing teachers, and *k* is the subscript indexing blocks, the model including treatment status, baseline test scores, one other student-level covariate, as well as a block indicator for the school and grade of the student's teacher can be written:

$$\text{(D. 1)}\ \gamma ij = \beta_0 + f(\beta_1, Z_{ij}) + \beta_2 X_{ij} + \beta_3 T_j + T_j f(\beta_4, Z_{ij}) + \sum_{k=1K} (\delta_k \beta_{k_j}) + y_j + u_{ij}$$

In the model, the outcome of student *i* of teacher *j* in block *k* depends systematically upon some function of the student's pretest score (*Z*), some other student characteristic (*X*), the treatment status of the student's teacher or classroom (*T*), a fixed effect of the block (*B*) to which the teacher belongs (δ), and random errors representing a random teacher effect (γ) and a random student-level error term (*u*). The error terms were each assumed to be independent and identically distributed (i.i.d.) and independent from one another and from the independent variables in the model.

We chose to model the functional form of the relationship between the pretest score (*Z*) and the outcome (*y*) parametrically, with linear, squared, and cubic terms. We also included interactions between treatment status and the linear, squared, and cubic pretest terms.

For the Ed Tech study, in addition to the nonlinear function of the pretest score (*Z*), we included a student's age in the model. For the TFA study, we used the second pretest score as the covariate—in the data-generating model for the math posttest, we included the reading pretest as a covariate; in the reading model, we included the math pretest as a covariate.

Estimating equation (D.1) using the original experimental data provided estimates of all of the model coefficients (including all of the block effects) along with the estimated variances of γ and *u*.[63] For TFA, we also had an estimate of the correlation between the individual error term for the math posttest and the individual error term for the reading posttest. All estimates, reported in Table D.1, were later used in the simulation.

Step Two: Simulate New Baseline Characteristics. The next step in the simulation was to generate baseline characteristics for the new observations. For Ed Tech, the relevant baseline characteristics were the pretest score and age. For TFA, the baseline characteristics were the math and reading pretest scores. In simulating values of these variables for the new observations, we wanted to preserve the joint distribution of these baseline characteristics and the correlation between these baseline characteristics and any block level fixed effect.

[63] We estimated a single variance for the residual term *u*. Alternatively, we could have allowed the residual variance to vary by treatment status to reflect heterogeneous treatment effects.

Table D.1. Estimated Parameters for Simulations

	Ed Tech	TFA Math	TFA Reading
Treatment Status	0.89 (0.45)	2.59** (.89)	0.63 (0.78)
Pretest	0.84** (0.02)	0.54** (0.04)	0.46** (0.05)
Pretest * Treatment	-0.01 (0.02)	0.05 (0.06)	0.11 (0.06)
Pretest Squared[a]	0.07 (0.04)	0.39** (0.13)	0.32* (0.14)
Pretest Squared * Treatment[a]	-0.05 (0.05)	-0.10 (0.18)	0.05 (0.20)
Pretest Cubed[b]	-0.74** (0.13)	0.06 (0.31)	0.47 (0.53)
Pretest Cubed * Treatment[b]	0.14 (0.19)	-1.46** (0.50)	-1.48 (0.74)
Age	-1.61** (0.27)		
Other Pretest		0.17** (0.02)	0.19** (0.02)
Variance of Random Student Error	10.98	11.90	11.80
Variance of Random Teacher Effect	2.63	2.45	0.57
Correlation of Student Error Terms (TFA)		0.37	0.37
Sample Size	**7183**	**1565**	**1591**

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original Ed Tech study.

[a] Coefficients and standard errors on squared pretest terms are divided by 100 to improve readability.

[b] Coefficients and standard errors on cubed pretest terms are divided by 10,000 to improve readability.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

Conceptually, we could have either simulated additional observations from existing blocks or simulated entirely new blocks of observations. We chose to simulate additional observations from existing blocks so that we would have estimates of the block fixed effects and be able to capture the correlation between the block fixed effects and the underlying test score distribution. We randomly assigned each observation to one of the existing blocks. Conceptually, this was like simulating new classrooms containing new student sample members within the grades and schools in which random assignment was already taking place.

For each block, we used the original data to calculate block-specific distributions of the baseline characteristics. In the case of TFA, for example, we calculated the mean math pretest score, the mean reading pretest score, and the correlation between the two pretest scores separately for each TFA block. New baseline covariate values were drawn from a block-specific, bivariate normal distribution. To be consistent with the initial test score distribution, we censored the simulated pretest values at the test score floors.

Step Three: Calculate Posttest Scores. Calculating the posttest required baseline covariates, a block-level fixed effect, a teacher random effect, and student level random error term. In the previous step, we randomly assigned new students to an existing block, transferring the estimated block effect from the experimental data, and generated the baseline covariates. Within blocks, students were randomly assigned to teachers. The number of teachers added to each block was determined by the number of new students assigned to the block and the average class size in the original sample. Each teacher received a random effect drawn from a mean zero normal distribution. The variance of the teacher random effect distribution was based on the estimated variance from the initial experimental model described above. The final component was the student level error term. This individual level error term was drawn from a mean-zero normal distribution using the calculated variance from the experimental model. For TFA, the individual error terms were drawn from a bivariate normal that allowed for correlation between the individual math error term and the individual reading error term.

With the parameter estimates from the experimental model and the generated covariates and error terms, we generated posttest scores for our simulated students. The final step was to censor the simulated posttest scores to preserve the test score floor in the initial sample. While the test scores also had a hypothetical ceiling, this was not observed in our data.

B. Evaluating the Simulated Data

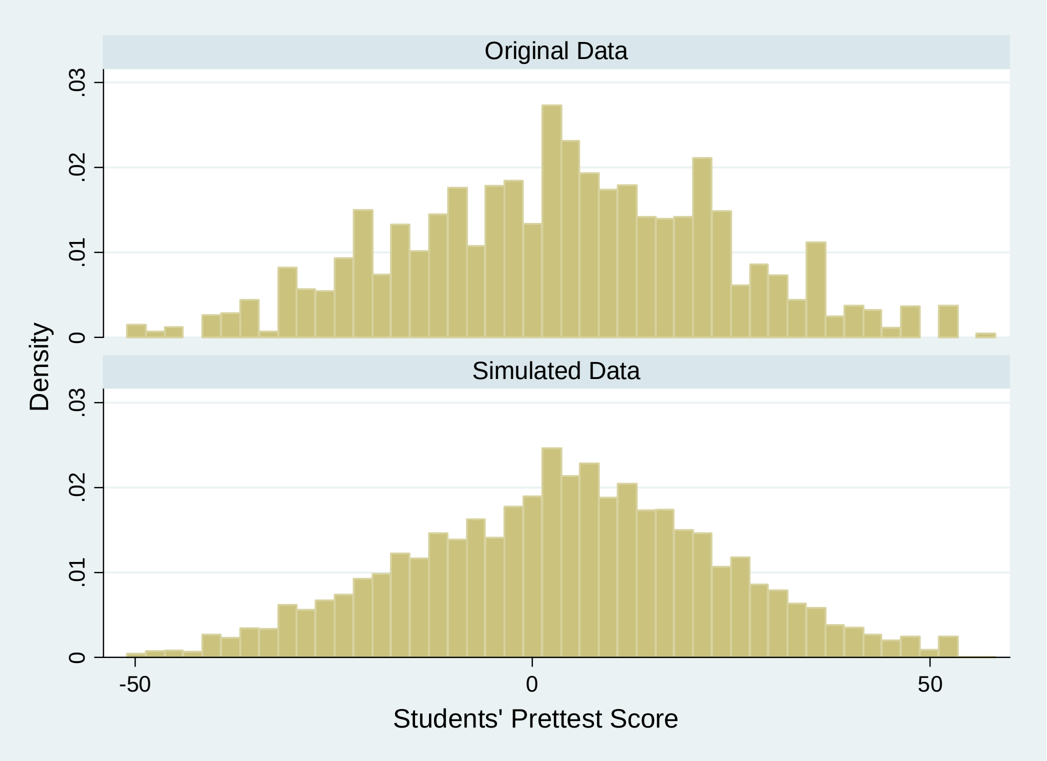
The simulation process aimed to preserve the joint test score distributions. This section presents evidence to validate the simulation. Figure D.1 presents the distribution of students' pretest scores for the Ed Tech original and simulated samples. One important difference between the two distributions is that the original test score distribution was much lumpier. The primary cause of the lumpiness in the original distribution was the limited possible values for tests scores. In the case of TFA, the small sample size also contributed to the lumpiness. Although we censored the simulated test score distributions at the test floor, we did not attempt to mimic the clumping of students throughout the test score distribution. While the simulated distribution was smoother, the general shape of the two distributions was similar.

Table D.2 compares the Ed Tech original and simulated data at different points in the test score distribution. In the original data, the 25th percentile of the recentered student pretest distribution was -14.5. In the simulated data, the 25th percentile was -12.8. The original and simulated samples were also similar at the 50th and 75th percentile. The standard deviation of the student pretest was

20.2 in the original sample and 19.6 in the simulated sample. In addition to preserving the test score distribution, the simulation aimed to preserve the correlation between variables. For Ed Tech, the correlation between the pretest and the posttest was 0.79 in both the original and the simulated samples.

Figures D.2 and D.3 show the student pretest distributions for the TFA math pretest and reading pretest. Again with TFA, the original data were much lumpier than the simulated data. The censoring of simulated test scores at the test score floor is quite evident in these figures. Although the actual test score floor was zero, the censored scores appear at different places in the distribution because the test scores were recentered based on grade-specific medians. The correlation between the math pretest and the math posttest was 0.60 in the original sample and 0.56 in the simulated sample. For reading, the correlation was 0.63 in the original sample and 0.56 in the simulated sample. The simulation also preserved the correlation between the math and reading pretest. The correlation was 0.52 in the original sample and 0.53 in the simulated sample.

Figure D.1. Histogram of Students' Pretest Scores for Original and Simulated Data: Ed Tech RD High Sample



Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Table D.2. Comparison of Original and Simulated Data: Ed Tech

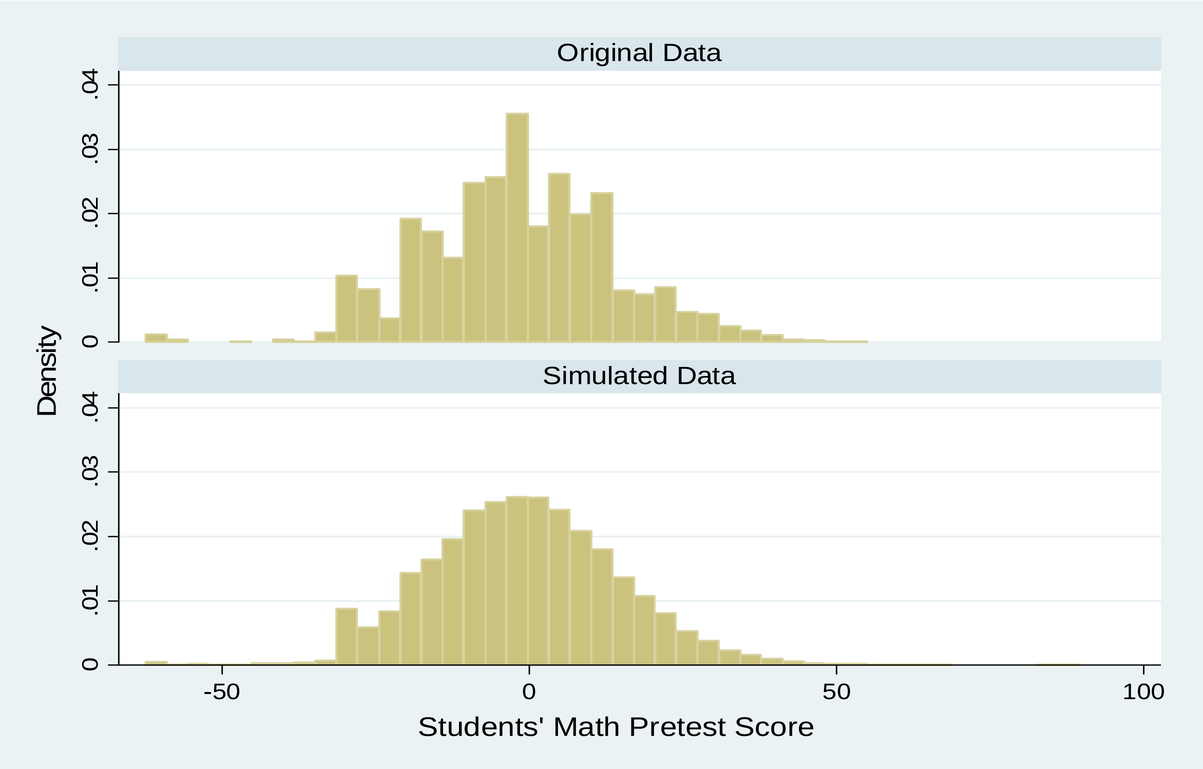
	Original Data	Simulated Data
Students' Pretest Score [a]		
25th Percentile	-14.5	-12.75
50th Percentile	0	0.5
75th Percentile	14.3	14.3
Standard Deviation	20.2	19.6
Students' Posttest Score [a]		
25th Percentile	-14.0	-12.0
50th Percentile	0	1.3
75th Percentile	13.3	14.8
Standard Deviation	19.8	19.0
Correlation of Pretest and Posttest	0.79	0.79
Sample Size	**7569**	**83921**

Source: Data from the Educational Technology Study (Dynarski et al. 2007).

Note: Observations are weighted to account for nonresponse and unequal probability of assignment to treatment. The table reports summary statistics for the sample used in the RD Replication study, not the original Ed Tech study.

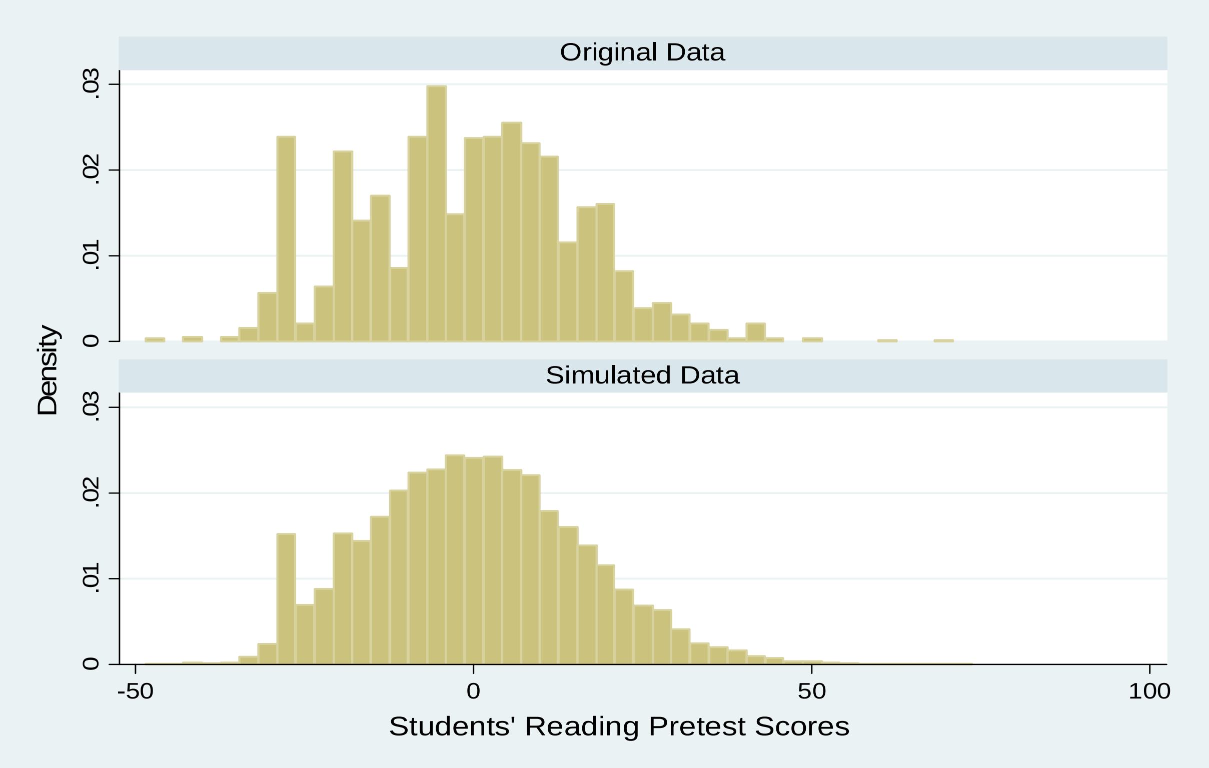
[a] Test scores are reported in re-centered NCE units. Re-centered test scores are calculated by subtracting the grade-specific median from the original test score.

Figure D.2. Histogram of Students' Pretest Scores for Original and Simulated Data: TFA Math RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

Figure D.3. Histogram of Students' Pretest Scores for Original and Simulated Data: TFA Reading RD High Sample



Source: Data from the Teach for America Study (Decker et al. 2004).

A comparison of the estimated impacts of these two interventions using an RD design based on the simulated data with the estimated impact using the original data, shown in Table D.3, supports the conclusion that the two methods produced results that are not significantly different from one another.

Table D.3. Original Regression Discontinuity (RD) Impact Estimates and RD Impact Estimates Based on Simulated Data by Data Set and RD Estimation Approach

	RD Original Data	RD Simulated Data
Ed Tech		
Local Linear	-0.06 (0.12)	-0.01 (0.02)
Sample Size	1513	19251
Parametric	0.00 (0.05)	-0.01 (0.01)
Sample Size	3681	46118
TFA—Math		
Local Linear	0.05 (0.23)	0.16** (0.04)
Sample Size	518	3262
Parametric	0.07 (0.17)	0.14** (0.04)
Sample Size	804	9730
TFA—Reading		
Local Linear	0.05 (0.20)	0.03 (0.04)
Sample Size	511	5906
Parametric	0.03 (0.11)	0.04 (0.03)
Sample Size	862	9776

Source: Data from the Educational Technology Study (Dynarski et al. 2007) and the Teach for America Study (Decker et al. 2004).

Note: Original RD estimates based on models presented in Chapters 3 and 4. The RD models with simulated data include the original data as well as the simulated data. Standard errors are shown in parentheses. All estimates and standard errors are shown in effect size units. Effect sizes are calculated by dividing by the standard deviation of the outcome variable among the full sample.

*Significantly different from zero at the .05 level, two-tailed test.
**Significantly different from zero at the .01 level, two-tailed test.

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA