A Randomized Experiment Comparing Random to Cutoff-Based Assignment

William R. Shadish  and Rodolfo Galindo
University of California, Merced

Vivian C. Wong, Peter M. Steiner and Thomas D. Cook
Northwestern University

Author Note

William R. Shadish and Rodolfo Galindo, School of Social Sciences, Humanities and Arts, University of California Merced; Vivian C. Wong, Peter M. Steiner and Thomas D. Cook, Institute for Policy Research, Northwestern University.

This research was supported in part by grant R305U070003 from the Institute for Educational Sciences, U.S. Department of Education. The authors thank Anja Göritz, at the University of Würzburg, Germany, for advice about web-based programming of this experiment.

Correspondence concerning this article should be addressed to William R. Shadish, School of Social Sciences, Humanities and Arts, University of California, Merced, 5200 North Lake Rd, Merced CA 95343. email: wshadish@ucmerced.edu
Abstract

Regression discontinuity designs (RDD) assign participants to conditions using a cutoff score, with those above the cutoff going to one condition and those below to another. Statistical theory shows that a correctly implemented and analyzed RDD gives unbiased effect estimates, just as in a randomized experiment. This study tests that theory by randomly assigning 588 participants to be in a randomized experiment or a regression discontinuity design, and then comparing results. Participants were tested on vocabulary and mathematics pretest and outcomes, and on many covariates, and were given a mathematics or vocabulary intervention. Results suggest that estimates from regression discontinuity designs approximate the results of randomized experiments reasonably well, but also raise the issue of what constitutes agreement between the two estimates.

*Keywords*: regression discontinuity design, randomized experiment, cutoff-based assignment.
A Randomized Experiment Comparing Random to Cutoff-Based Assignment

A research industry has blossomed in the last two decades aimed at identifying effective treatments. It occurs under labels like evidence-based practice, within organizations like the Institute of Education Science’s What Works Clearinghouse, or in response to third-party payer requests for evidence that they are reimbursing effective services. It particularly affects researchers in education, public health, medicine, clinical psychology and similar disciplines where professionals intervene to help others. And it has led to renewed interests in all variety of research methods for investigating effective treatments, but especially experimental and quasi-experimental designs (Shadish & Cook, 2009). For both theoretical and practical reasons, randomized experiments (REs) remain the most often preferred methodology for assessing treatment effects. REs assign participants to two or more groups by chance to assess the effects of treatment (Shadish, Cook & Campbell, 2002). Random assignment allows effect estimates that are unbiased, that is, where the expectation of the effect equals the effect in the population.

Close behind in preference, however, is a much less well-known methodology that has been increasingly used in the last 15 years, the regression discontinuity design (RDD, sometimes called a cutoff-based experiment; Cook, 2007). Invented by Thistlewaite and Campbell (1960), RDDs assign participants to conditions using a cutoff score on an observed assignment variable (Shadish et al., 2002). Consider the following example. When prison inmates are first released, they often lack a job and other financial resources to help them become productive members of society. Do some of them return to crime after leaving prison in order to get those resources? Will providing them with funds on their release reduce future offending? Berk and Rauma (1983; Rauma & Berk, 1987) tried to answer the last question when the State of California passed legislation giving unemployment compensation to newly released prisoners, but only if they had
worked more than 652 hours over the previous 12 months while in prison. Those who worked fewer hours were ineligible. Berk and Rauma found that those receiving unemployment compensation had a recidivism rate 13% lower than controls. The literature now contains an increasing number of examples like this one of RDDs in education, economics, medicine, and other fields.

But does the RDD yield an accurate effect estimate? In theory it is known to yield an unbiased estimate, just like the RE. The reason is essentially that because the assignment variable is completely known and measured, the researcher can adjust for differences in selection to obtain an unbiased and consistent estimate of the treatment effect (Goldberger, 1972, 2008; Rubin, 1977). More intuitively, one can think of RDD as similar to a randomized experiment at the cutoff. The assumption here is that differences between units with assignment scores just above and below the cutoff (i.e., one or two points apart on different sides of the cutoff) are due almost entirely to chance (e.g., random measurement error). In theory, then, results from a RDD should especially well-approximate results from a parallel randomized experiment at the cutoff.

For several reasons, however, estimates from RDDs may not match those from REs as well as theory suggests. For example, just as is the case in a randomized experiment (RE), practical problems in the implementation and analysis of an RDD can compromise the effect estimate it yields. The most common implementation problems occur when assignment to condition does not adhere strictly to the cutoff, and when units manipulate their assignment scores to receive or avoid treatment. In both cases, the assignment mechanism into treatment conditions becomes partially obscured to the researcher, introducing the potential for selection bias. The central analytic problem is the need to correctly model the unknown functional form (shape) of the relationship between the assignment variable and the outcome variable near the
cutoff. Such modeling is not necessary in REs, trivially because REs do not have an assignment variable like that in RDD, nontrivially because members of the treatment and comparison conditions in an RE are equally distributed across the range of any variable so the effect can be computed no matter where it would be estimated on such a variable. The latter is not true in RDD where only members of one condition are left of the cutoff and only members of the other are right of the cutoff, so that estimating effects at the cutoff offers the best chance to use both treatment and comparison group members without heroic assumptions about what might have happened to them far from the cutoff where they do not exist. Recently, economists have used non-parametric techniques for estimating treatment effects at the cutoff (Hahn, Todd & van der Klauuw, 2001). These procedures relax the need to model the response function away from the cutoff, but introduce a bias versus efficiency tradeoff in selecting the correct bandwidth for the smoothing parameter. In addition, the RDD has at least two design limitations as compared to the RE. First, it has less statistical power than the experiment by a magnitude of two to five times (Goldberger, 1972, 2008; Bloom, Kemple, Gamse, & Jacob, 2005; Schochet, 2009). Second, the procedure identifies treatment effects well only at the cutoff, limiting the generalization of results if treatment effects across the range of the assignment variable are not constant, and making selection of a policy-relevant cutoff of special importance.

However, even if all these practical problems were solved, RDD still might yield a different answer from RE because the RE and the RDD only estimate the same causal estimands if treatment effects can be assumed to be constant over the range of the assignment variable. This is because RE and RDD provide two different estimates given how they are typically analyzed. The RE estimates the average treatment effect (ATE) and the RDD estimates the average
treatment effect at the cutoff (ATEC) of the assignment variable. These estimates coincide with constant treatment effects, but otherwise can diverge.

For all these reasons, researchers and policymakers are interested in empirical comparisons of RE and RDD show that they yield the same results. Several implementations of the regression discontinuity design have been compared to similar randomized experiments to test the comparability of their estimates (Aiken, West, Schwalm, Carroll, & Hsiung, 1998; Buddelmeyer & Skoufias, 2004; Black, Galdo, & Smith, 2007; Berk, Barnes, Ahlman, & Kurtz, 2010). These within-study comparisons take a causal estimate from an experiment and compare it to the estimate from a RDD that may share similar settings, interventions, and/or measures, but with different subjects. The goal of these studies is to assess whether the RDD produces the same causal estimate as the randomized experiment when implemented in the real world. Cook and Wong (in press) summarize results from the earlier within-study comparisons (Aiken et al., 1998; Buddelmeyer & Skoufias, 2003; Black et al., 2005) and we do so here, along with the most recent addition by Berk et al. (2010).

Aiken et al. (1998) used participants enrolled in a remedial, standard, and remedial/standard course from an incoming freshman class at a large state university. In the RDD, assignment of 141 students to the remedial class (or not) was based on an ACT or SAT cutoff score, capitalizing on the fact that students who score less than or equal to 16 on the ACT or 380 on the SAT must take remedial English composition. To increase the similarity of the RDD population to the population in the RE, participation in the RE was limited to 123 students whose ACT and SAT scores fell just below the tests’ respective cutoffs. Aiken et al. analyzed results with an ordinary least squares regression. Outcome measures included students’ performance on an essay, and on a multiple choice Test of Standard Written English (TSWE).
Aiken et al. found that the RDD yielded a pattern of results consistent with those obtained from the randomized experiment: Larger effect sizes were obtained for the TSWE than for the written essay for both assignment variables. Although results from the RDD and RE were comparable in terms of direction, they were not always identical in magnitude, and tests of significance were generally but not always in accord as to whether the effect size was reliably different from zero. Aiken et al. (1998) did not test the significance of the differences between effect sizes from the two designs, and they did only minimal modeling of functional form in their RDD.

Buddelmeyer and Skoufias (2004) evaluated PROGRESA, a poverty alleviation program in Mexico. Villages were randomly assigned to PROGRESA, and the experiment compared eligible families in villages randomly assigned to PROGRESA with eligible households in control villages. Eligibility depended on scoring below a cutoff on a family material resources scale, which enabled RDD within treatment villages. Here, the RE estimated the average treatment effect (ATE) for all eligible families in treatment and control villages, while the RDD estimated the average treatment effect at the cutoff (ATEC) in experimental treatment villages. Thus, the authors truncated the experimental sample to include only the subgroup of families with scores immediately below the RDD cutoff to approximate the ATEC estimate. Sample sizes for both design studies were large, about 10,000 cases in the RDD and RE, and the data were analyzed in several nonparametric regressions using kernel density functions, all predicated on weighting observations closer to the cutoff more heavily. Two outcomes, school attendance and participation in work-related activities, were examined for boys and girls over three rounds of data collection. Although the authors found that the RE and RDD yielded comparable estimates for two of three rounds on school attendance, and for all three rounds on participation in the workforce, they found discordant results for at least one round of school attendance—a pattern
that was replicated for both boys and girls. The experimental design showed small but significant
effects on boys’ and girls’ school attendance in round two, but the RDD showed no such effects,
with treatment estimates being close to zero. However, similar to Aiken et al. (1998),
Buddelmeyer and Skoufias (2004) did not test whether the effect estimates from the RDD were
significantly different from those from the RE.

Black et al. (2007) evaluated the worker profiling and reemployment services system
using a RDD in a job training experiment in Kentucky. In the RDD, individuals were assigned to
job training based on a cutoff score from a 140-item test predicting the likelihood of being long-
term unemployed and collecting unemployment benefits. The exact cutoff varied from week to
week depending on the number of treatment slots available in each of the many offices involved,
so RDD treatment effects were identified across multiple cutoff points. The sample included
4,465 claimants for the RDD estimates. The randomized experiment took over 1900 persons who
had tie scores at the RDD cutoffs, and randomized participants to treatment or control. Thus,
Black et al. was the only comparison to embed an experiment within the RDD, allowing the
authors to estimate an ATEC for the RE, and to do so from both “above” and “below” the cutoff
for the RDD. The dependent variables were weeks of unemployment insurance benefits, amount
of such benefits and annual earnings from work. Black et al. analyzed results using a variety of
parametric and nonparametric estimators, and they did test the statistical significance of the RE-
RDD difference. They found that that the local linear kernel regression showed the least bias,
though the multivariate parametric analyses also produced results that were comparable to the
experimental results. The simple Wald estimator yielded estimates that were similar to the RE
results in terms of direction of effects, but were generally more biased than estimates produced
by the local linear kernel and parametric approaches. As expected, most analyses suggested that bias was least when restricting the sample to participants closest to the cutoff.

Berk et al. (2010) examined whether reduced supervision of parolees and probationers deemed at “low risk” for public safety resulted in increased rates of arrest one year later. The authors used a forecasting model to assign 30,000 parolees and probationers in Philadelphia with “risk” scores for recidivism. Those who scored higher than .50 received standard supervision while those who scored less than .50 received reduced supervision. The RE consisted of 1559 “low risk” probationers and parolees who were randomly assigned to standard or reduced supervision. Thus, the RE sample consisted of offenders with lower risk scores than those included in the RDD sample. The authors used a generalized additive model (GAM) for analyzing treatment effects, which modeled the outcome as a flexible and smooth function of the assignment variable. Three outcomes were examined: (1) arrests for any new crimes; (2) arrests for serious crimes, and (3) arrests for crimes of violence. For all three outcomes, Berk et al. found that the RDD replicated the RE result in size, direction and statistical significance patterns, and that the RE estimates were not significantly different from the RDD estimates.

These studies highlight at least four challenges with using within-study comparisons to validate empirically the performance of the RDD. The first is to decide which statistical estimates should be used to compare RE and RDD. From a policy perspective, decision makers may want to know whether the RDD ATEC can replicate the same policy conclusion from an RE ATE because that is how the two designs are typically analyzed. From a methodological perspective, researchers may want to know whether the RDD yields valid ATEC results despite implementation and analytical challenges in the field. Three of the above within-study comparisons (Aiken et al. 1998; Buddelmeyer & Skoufias, 2004; Black et al., 2007) limited the
experimental samples to observations within a narrow range below or above the RDD cutoffs, generating an approximation to the ATEC as the benchmark. Berk et al. (2010) compared the RE ATE for the full range of low risk offenders with the RDD ATEC at the cutoff, comparing two different causal quantities if the treatment effect interacts with the assignment variable.

A second challenge is that the RDD and RE contrast may be correlated with other variables related to the study outcome. Most studies reviewed here were measured in the same ways at similar time frames, but sample differences between the RDD and RE may have remained. For example, in Aiken et al. (1998), participants in the RDD sample had to take the course that their SAT or ACT assigned to them while those in the RE could refuse randomization into treatment conditions. Thus the difference in how participants were assigned to conditions in RE versus RDD is potentially confounded in the Aiken et al. case with differences in the kinds of students in RE versus RDD on variables like motivation to refuse participation. This kind of confound is common in nearly all the previous studies comparing RDD to RE. Third, the early within-study comparisons lacked consensus on a standard criterion for achieving correspondence, raising the question of “how close is close enough?” (Wilde & Hollister, 2007). Aiken et al. (1998) and Buddelmeyer and Skoufias (2004) looked at the size and direction of RDD and RE effects, and patterns of statistical significance, but Black, Galdo, and Smith (2007) and Berk et al. (2010) also conducted tests of statistical difference between RDD and RE results. Finally, none of the studies reviewed here discussed a protocol for “blinded analyses” prior to comparing RDD and RE results. The issue here is that an RDD analyst encounters many decision points in which he or she may deliberately or inadvertently skew the analysis to increase the likelihood of replicating already known benchmark results from the RE. Because of this concern, Cook, Shadish, and Wong (2008) recommend that separate teams of researchers analyze the non-
experimental and experimental data, with both sides blinded to the analytic procedures and results of the other.

Despite these issues, the four studies suggest a number of lessons about the performance of the RDD in the field. First, they generally support the hypothesis that the RDD produces similar causal estimates to the RE across the majority of comparisons attempted. Second, whether tested by statistical or practical significance, a nontrivial percent of the comparisons did not yield the same results. Third, the three studies used quite different statistical methods to identify the RDD effect estimate, with different methods producing different results.

Given the latter two lessons, questions remain about whether the results of RDD match those from RE (at the cutoff) as well as statistical theory says they should. The present study revisits this issue, adding to the literature in a unique way by using a highly controlled laboratory approach to the comparison. Specifically, we adapt a laboratory analogue paradigm that was previously used to randomly assign participants to be in a randomized experiment or in a nonrandomized experiment in which they could choose their training, and where they were otherwise treated identically (Luellen, Shadish, & Clark, 2005; Shadish, Clark & Steiner, 2008; Shadish, Luellen, & Clark, 2006). The current study adapts that design by randomly assigning participants into either a randomized experiment or a regression discontinuity design in which they are otherwise treated identically and simultaneously. The latter random assignment makes the present study unique, ensuring that both the RE and RDD test treatments on the same populations, and making more plausible comparisons between RE and RDD that assume they share the same functional form. Further, identical and simultaneous treatment and assessment of participants ensures that neither time nor conditions are confounded with the RE-RDD contrast.
Finally, the controlled conditions vastly reduce problems that might otherwise occur with attrition and partial treatment implementation.

**Method**

**Participants**

This study used UC Merced Human Subjects Pool students (N = 588). The mean age was 19.04 years (SD =1.42). The gender break down was 343 females (58%) and 245 males (42%). Ethnicity included 31.1% Hispanic, 30.3% Asian, 23.3% White, 6.0% Black, Pacific Islander 2.0%, and 7.0% other. Majors were 33.0% math intensive, 50.5% not math intensive, and 15.1% undecided, and 1.4% unknown. Large numbers of undergraduates from all over the university take introductory psychology as an elective, and only 29.4% of the participants were Psychology majors.

Nineteen (3.2%) of the original 588 participants, eight (4.1%) from the RE and eleven (2.8%) from the RDD, failed to take both the mathematics and the vocabulary posttests, and were excluded from remaining analyses. Of the eight in the RE, five were from the vocabulary condition and three were from the mathematics condition. Of the eleven from the RDD, nine were part of the vocabulary condition and two were from mathematics. In addition, two data points were missing for marriage status, one from a 20 year old participant and the other an 18 year old participant. We imputed their status as “single” based on the likelihood of this being the case, as it was for 98.8% of our sample. All other variables had 100% complete data. Due to the small number of missing posttest values we ran our main analyses without imputations. A secondary analysis with regression-imputed posttest values showed that the estimated treatment effects are almost identical to those without imputation.

**Procedure**
This was a web-based experiment in which participants from the UC Merced subject pool followed a link online to a secure server to participate (Göritz & Birnbaum, 2005). After informed consent, the study began with baseline tests that measured participants’ vocabulary and mathematics skills. Then $N = 391$ of the participants were randomly assigned to the regression discontinuity design, and $N = 197$ to the randomized experiment (Figure 1). More participants were assigned to the RDD in order to improve its statistical power, which is lower than the power of a randomized experiment. Those assigned to the randomized experiment were randomly assigned to mathematics or vocabulary training. Those who were assigned to the regression discontinuity design were assigned to training based on a cut-off score of 20 on the vocabulary pretest. Those scoring 20 or above received mathematics training, those scoring below 20 received vocabulary training. Twenty was used as the cutoff because that was the mean of the vocabulary pretest in Shadish et al. (2008). The training sessions in both experiments were identical. After training, all participants were assessed on both mathematics and vocabulary outcomes. This design ensured that all participants were treated identically in all respects except for assignment method.

**Pretests**

Instructions appeared on a computer screen and were scored by a computer program for all of the following pretests: (1) Demographics Questionnaire, prepared by the present researchers, gathered data about participant age, sex, ethnicity, marital status, major, high school algebra enrollment, high school calculus enrollment, college algebra enrollment, college calculus enrollment, attitude on English courses, views on mathematics courses, preference for mathematics or vocabulary, willingness to take math electives, willingness to take English electives, future training interest, and major area of study (Math or Language intensive); (2) the
Vocabulary Test (Educational Testing Services, 1962) measured vocabulary skills; and (3) the Arithmetic Aptitude Test (Educational Testing Services, 1993) measured mathematics skills.

Treatments

A series of web pages presented interventions to teach either 50 advanced vocabulary terms or five algebraic concepts. The vocabulary training included a novel term, its phonetic spelling, a sentence in which the word was used, and a request to the participant to use the word in a sentence he or she typed on the webpage. The mathematics training included five rules for transforming exponential equations, several examples in which those rules were applied to algebraic formulas, and a practice test in which two questions for each rule were given. We compared two treatment conditions rather than comparing treatment to no treatment in order to create two tests of the RE-RDD difference: one for the effects of vocabulary training on vocabulary outcome and one for the effects of mathematics training on mathematics outcome.

Posttest

A 50-item posttest contained 30 vocabulary items (15 presented in training and 15 new) and 20 mathematics items (10 presented and 10 new), presenting vocabulary first and mathematics second for all participants. This posttest was given to all participants regardless of training.

Implementation Problems

Minor problems emerged in the web-based implementation of the present experiment that we learned only after it ended. For example, participants could use the back button on their web browser to be reassigned to conditions until they got the condition they wanted. Similarly, because participants mostly completed the experiment on their personal computer and not in a laboratory, they may have used online dictionaries or scientific calculators. Time limits on tests
minimized the latter problem; and we also surveyed participants retrospectively asking if they used such aids, and virtually all said no. Participants taking the experiment at home may have been distracted from the experiment or otherwise stop participating in it for a period of time, and then may have been timed-out during administration of the timed outcome measures. Though we have no reason to think these problems could account for the exact pattern of our results, nonetheless it would be preferable to run this experiment in a more controlled setting, though that would make large sample sizes harder to achieve.

**Analyses**

The procedure to analyze the RDD and RE data consisted of two stages. In the first stage, RE and RDD were analyzed independently by two groups of authors. The first and second authors analyzed the RE, and the other authors analyzed the RDD, each without knowledge of the others’ analytic strategy and results, except to ensure that the same covariates were used in both analyses. Teams also did not confer in advance which causal estimands (ATE versus ATEC) to compare. This was to ensure that knowledge of the outcome from one analysis would not bias how the other analysis was conducted. The independent analyses of the RE and RDD were set up to mimic what researchers would do in practice when they have RE or RDD data only. Thus, the initial analysis focused on a question of direct policy interest: Do RE and RDD result in different effect estimates when they are independently analyzed according to the state of art?

In the second stage of analyses, our goal was to examine the comparability of RE and RDD results by (1) focusing on the same causal estimand and (2) holding the analytic methods constant by using the same type of analysis for the RE and RDD. The research question addressed by the second stage exploratory analysis is of more theoretical interest: Do RDD and
RE result in comparable effect estimates when both are evaluated at the cutoff (ATEC) with the same methods?

**Independent Analysis.** Both teams of analysts began by examining whether the RE and RDD were implemented properly. The RDD analysts found no evidence of treatment misallocation or manipulation of the assignment mechanism by participants. The RE analysts found no significant differences in covariates between treatment and control groups, indicating that the randomization procedure worked. Both teams of researchers included in addition to the vocabulary pretest (the assignment variable) all 14 baseline covariates in RE and RDD analyses, which included controls for the mathematics pretest score, age, ethnicity, past and current academic interests and achievements, and whether their major was mathematics intensive or not (Table 4 lists all predictors, most of them being dummy variables).

Next, the RDD analysts explored the functional form of the RDD data with various parametric models. They investigated the appropriateness of linear regression models including the linear, quadratic and cubic polynomial of the centered assignment variable, as given by equations (2)-(4) in Table 1. To avoid strong functional form assumptions—including the assumption of constant treatment effects—they interacted the polynomial with the treatment indicator, thereby allowing full flexibility for the polynomial to the left and right of the cutoff. This type of modeling also implies that the RDD analysts tried to estimate ATEC at the cutoff. Although the analysts examined the RDD data using non-parametric approaches (local linear regression, equation (6) in Table 1), they chose the best fitting parametric model for both outcomes for two reasons. First, residual plots indicated that parametric models reasonably represented the functional form of the RDD data. Second, the medium sample size resulted in large bandwidths for the nonparametric smoothing parameter and comparatively large standard
errors of the treatment effects. Using Akaike’s information criterion (AIC) and $F$-tests, the RDD analysts then selected the best fitting parametric model for both outcomes. For the effects of vocabulary training on the vocabulary outcome, they chose the quadratic model (Table 1, equation 3), for the effects of mathematics training on the math outcome they selected the linear model (equation 2).

Masked to RDD analysts’ decisions and results, the RE analysts estimated a covariance adjusted average treatment effect for the overall RE population, which is the causal estimand usually estimated in a randomized experiment (Table 1, equation 1). All covariates listed in Table 4 were included as linear predictors.

**Exploratory Analysis.** The preceding RDD and RE analyses focused on different causal quantities: ATEC in RDD and ATE in RE. Since ATEC and ATE may differ unless the assumption of constant treatment effects is met, we conducted further exploratory analyses that used the same analytic methods for both the RDD and RE data, and that forced the two designs to estimate the same causal quantity, the ATEC at the RDD cutoff, with the same functional form restrictions. In the latter case, even if we cannot know for certain which functional form is the correct one, we can reasonably assume—by virtue of the original random assignment of participants to RE or RDD—that the functional form should be identical within sampling error for both designs.

Table 1 shows all the models that we estimated: Models (2)-(4) represent parametric models with a linear, quadratic and cubic polynomial of the centered assignment variable that is interacted with treatment. Moreover we estimated treatment effects using a semi-parametric approach, particularly a generalized additive model (GAM, Hastie & Tibshirani, 1990) as given by model (5). Here, the outcome of each treatment group is modeled as a smooth function of the
centered assignment variable via a knot free thin plate regression spline basis (Wood, 2006). The smoothing parameter that controls the penalty for the ‘wiggliness’ is determined by minimizing the generalized cross validation score. Although the RDD analysts rejected a nonparametric estimation of the treatment effect in the independent analysis, we present results from local linear regression analyses since nonparametric approaches are often used to estimate ATEC in RDD (Imbens & Lemieux, 2008). Table 1 gives the details on our local linear regression model, which was implemented using a triangular kernel with an optimal bandwidth choice as suggested by Imbens and Kalyanaraman (2010). In order to increase the comparability of the local linear regression estimates for the RE and RDD we determined the optimal bandwidth for the RDD but used the same bandwidth for the RE. Due to the smaller RE sample size, this results in rather large standard errors for nonparametric RE estimates. However, with an optimal bandwidth selection for the RE data we would have obtained reduced standard errors but no significant changes in results. We estimated standard errors for the local linear regression estimates from 1,000 bootstrap samples.

Correspondence between RDD and RE results was assessed in three different ways by examining the size and direction of effects, patterns of statistical significance, and direct tests of statistical difference between RDD and RE effects. These direct tests are $t$-tests comparing regression coefficients from two independent samples. The $t$-test statistic is given by

$$t = \frac{\hat{\tau}_{RDD} - \hat{\tau}_{RE}}{SE_{RDD-RE}}$$

where $\hat{\tau}_{RDD}$ and $\hat{\tau}_{RE}$ are the estimated treatment effects from the RDD and RE, respectively, and $SE_{RDD-RE} = \sqrt{SE_{RDD}^2 + SE_{RE}^2}$ is the estimated standard error of the difference in treatment effects with $SE_{RDD}$ and $SE_{RE}$ the corresponding regression standard errors.

Results
Table 2 shows RDD and RE effects of the mathematics training on the mathematics outcome, and Table 3 shows RDD and RE effects of the vocabulary training on the vocabulary outcome. For both sets of outcomes, we present results first from our independent analyses, which compare RE ATE with RDD ATEC results, and then from our exploratory analyses, which compare RE and RDD ATEC results using parametric, semi-, and nonparametric approaches. Models (2) through (6) of Tables 2 and 3 show changes in effect estimates as the analytic approach become more flexible, and treatment effect estimates become more “local” to the RDD cutoff.

**Independent Analyses**

For the effects of mathematics training on the mathematics outcome, the randomized experiment yielded an average treatment effect of 2.53 points (Table 2) while the RDD yielded a average effect at the cutoff of 1.87 points. Despite the RE and RDD estimating different causal quantities, both designs yielded effect estimates that were comparable in terms of direction, size, and statistical significance patterns. The effect size difference between the two estimates was small (-0.15 SD), and the *t*-test indicated no reliable differences between the RDD and RE results (*p* = 0.43). For the effects of the vocabulary training on the vocabulary outcome (Table 3), the RE and RDD also yielded comparable results, albeit not as close as the estimates for the mathematics training. The randomized experiment yielded an average treatment effect of 4.46 points (Table 3), and the RDD yielded an ATEC of 5.95. Both the RE and RDD produced treatment effects that were large, positive, and statistically significant, but the difference between the two results approached medium size (0.41 SD). A *t*-test of the difference between the RE-ATE and RDD-ATEC estimates was not significant (*p*-value: 0.10).
Note, that the results would have been the same if the RE analysts would have estimated the experimental treatment effect without any covariate adjustment. For the mathematics outcome the RE treatment effect would have been 2.33 points with a standard error of .64 points; for the vocabulary outcome the RE effect would have been 4.44 points with a standard error of 0.52 points.

**Exploratory Analyses**

Next is a series of exploratory analyses that compared RE ATEC estimates with RDD ATEC results. For the effects of the mathematics training on the mathematics outcome, we found high correspondence in RE and RDD results. Figures 2 and 3 provide graphical depictions of the RDD and RE response functions, respectively. They suggest that the linear model appears to be a reasonable specification of the RDD and RE response functions. The robustness in estimates from the parametric, semi- and non-parametric analyses also confirm this conclusion. For the linear parametric model, the RE treatment effect was 2.64 and the RDD effect was 1.87 points; for the quadratic model, the RE effect was 2.70 and the RDD effect 1.88 points; and for the cubic model, the RE effect was 2.96 and the RDD effect 2.64 points. The effect size differences ranged from -0.17 SD for the linear model to -0.07 SD for the cubic model, and there were no statistically significant differences between RDD and RE results. When we relaxed the functional form assumption further by using GAM to estimate treatment effects, the RE produced a result of 3.24 and the RDD a result of 1.87 points, not significantly different. Using the nonparametric local linear regression approach, the difference between the RE treatment effect of 2.69 points and the RDD effect of 1.98 points was insignificant, too. Note that the standard errors of the treatment effects tend to increase as the flexibility of the regression models
increase. The RE’s large standard error for the local linear regression estimate is mainly due to the narrow bandwidth which we determined using the larger RDD sample. Also note the lower efficiency of parametric RDD estimates in comparison to corresponding RE estimates. Despite a two times larger RDD sample, RDD standard errors of treatment effects are larger than corresponding RE standard errors.

For the effect of vocabulary training on the vocabulary outcome, the pattern of correspondence in the RE and RDD results was somewhat less convincing. Using a parametric linear functional form, the RE produced a treatment effect of 4.35 and the RDD an effect of 5.10 points. Both estimates were large, positive, and significant, and the effect size difference was small (0.21 SD), with no significant differences between the RE and RDD results. However, graphical inspections (see Figures 4 and 5) and residual analyses rejected the linear function as an appropriate specification for the RE and RDD data, so we looked more closely at treatment effect estimates produced by more flexible parametric response functions (i.e. quadratic and cubic models), and by a semi- and nonparametric approach. For the quadratic function, the RE produced a treatment effect of 3.92 and the RDD an effect of 5.95 points. Although the direction and statistical significance pattern of this result was comparable, the effect size difference was 0.56 SD and the null hypothesis that the RE-ATEC equaled the RDD-ATEC was rejected ($p = 0.04$). The parametric cubic function produced an RE treatment effect of 3.64 and an RDD effect of 6.16 points. The effect size difference was larger than for the quadratic model (0.69 SD), and again, the null hypothesis was rejected ($p = 0.04$). The GAM approach yielded an RE effect of 2.74 and an RDD effect of 5.81 points. The effect size difference was also 0.84 SD, and again the null hypothesis was rejected ($p = 0.01$). We did not obtain a significant difference for the
local linear regression estimates (.63 SD) because of the large standard error for RE (with an optimal bandwidth choice for RE the difference would have been significant).

A possible reason for these significant differences might be that the initial random assignment procedure to the RE and RDD failed to result in covariate balance over those conditions. For each covariate, we compared the RDD and RE overall mean scores but also the mean scores within a narrow interval (± 3 points) around the RDD cutoff (Table 4). Although there were no significant differences between the RE and RDD means for 13 of the 14 baseline covariates at the cutoff, the overall RDD mean for the math pretest score was significantly larger than the RE mean by 0.72 points (0.26 SDs). The mean difference was also of the same magnitude close to the assignment variable cutoff. This implies that the RE and RDD ATEC estimates refer to different populations regarding the math pretest. Hence we conducted an additional set of exploratory analyses to adjust for this difference. Using RDD data, we first estimated the average math pretest score at the RDD cutoff using a local linear regression of the math pretest on the assignment variable, with the estimate being 6.77 points. Then we centered the math pretest variables in both the RE and RDD at 6.77. We then redid the previous RE and RDD analyses but this time adding the newly centered math pretest score and its interaction with the treatment indicator. The results are in the last section of Tables 2 and 3. These analyses reduced the size of the difference between the RE and RDD estimates in all cases, and none of those differences were statistically significant at $p < 0.05$, though three were different at $0.05 < p < 0.10$ for vocabulary.

**Discussion**

In the present study, a regression discontinuity design well-approximated the results of a randomized experiment in most but not all respects. First, and most pertinent to policy, the
comparison of RE-ATE to RDD-Atec suggests that a regression discontinuity design as it is typically analyzed yielded about the same effect size estimate as a randomized experiment as it is typically analyzed—even though the analyses estimated different parameters for the two designs. This is a cautious conclusion because one can imagine a RDD-Atec that differs substantially from an RE-ATE in the presence of treatment interactions and certain kinds of nonlinear functional forms. Second, the analyses that constrain the RE and RDD to estimate the same parameter show that they give the same answer when similarity is judged by the size, direction, and statistical significance of the treatment effect, but less consistently when similarity is judged by the statistical significance of the difference between the RE-Atec and RDD-Atec estimates. Of course, this second set of results is of little policy relevance because few analysts would estimate such a RE-Atec in practice, but it might constitute one appropriate test of the statistical theory. In addition, these results are consistent with four previous studies (Aiken et al., 1998; Berk et al., 2010; Black et al., 2005; Budelmeyer & Skoufias, 2003), but with a stronger methodology—the participants in the RDD were identical to those in the RE except for sampling error, they were treated identically and simultaneously, and their data should share a common functional form. This reduces concern about potential systematic confounds with assignment methods that might have been present in at least some of those past studies.

Understanding Possible Discrepancies

What should be made of the finding that the RE-Atec estimates differed significantly from the RDD-Atec estimates for the vocabulary outcome in three of the eight exploratory analyses, and that the difference is larger than a half standard deviation? The four prior studies that compared these two designs did not find perfect agreement between them either. For example, while Black et al. (2005) obtained excellent agreement when comparing their full
experimental estimates to a parametric RDD-ATEC, other comparisons were not as concordant. Their comparison of RE to RDD with a local Wald estimator yielded significant differences on 10 of 24 tests; with a smoothed local Wald estimator, 5 of 24 were significant; with a Hahn-Todd-van der Klauuw kernel estimator, 8 of 24 were significant; and other analyses yielded similarly variable results. This led Black et al. to conclude that the results were “sensitive to changes in the non-experimental samples used to estimate the counterfactuals and the outcomes of interest” (p. 44). So it may be that some part of the differences is due to exactly how the data are analyzed.

We also considered whether something about the vocabulary training made it inherently less amenable to estimation, but deemed that unlikely for one reason. Shadish et al. (2008) used almost identical treatment conditions with nearly identical predictors and outcome measures in a nearly identical design comparing random assignment to self-selection into conditions. They found that the adjusted nonrandomized experimental results for the vocabulary condition better approximated results from the randomized experiment compared to the mathematics condition, opposite of the present study. To explain this, one would have to postulate an interaction between the substantive nature of the treatment condition and the type of design, and then find a substantive explanation for why that might be the case. Compared to the alternative explanations, this seems implausible.

A more likely explanation for the lower consistency between RE-ATEC and RDD-ATEC for vocabulary may be sampling error in the original random assignment to either random or cutoff-based assignment at the start of the experiment. This explanation garners some support from the fact that the exploratory analyses that included an interaction between math pretest and assignment condition reduced the vocabulary differences between RE-ATEC and RDD-ATEC.
from significance to $0.05 < p < 0.10$. Rubin’s (2008) commentary on Shadish et al. (2008), which appeared after the data in the present study were gathered, argued for the need to ensure both that the original random assignment to assignment method resulted in covariate balance over assignment methods, and that the same was true for covariate balance after the random assignment of participants to vocabulary or mathematics training. Either imbalance would call into question how well the randomized experiment serves as a gold standard for the nonrandomized results. If one had the full participant pool before the start of the experiment, then matching participants on pretest covariates prior to both random assignment would probably improve balance considerably. That would not work in the present design where participants enroll in the study over time. More feasible is pre-assignment stratification on important covariates, where the strata are defined based on previous studies, for example, based on the present study and on Shadish et al. (2008) if the aim was to replicate or extend results from the present methods.

**Randomized Experiments as Gold Standards**

Taking this one step further, the present study leads to questions about what it means to claim that estimates from the randomized experiment are the gold standard. This term can imply either that the randomized experiment is generally the best design for causal inference, or that the failure of a nonrandomized estimate to match the randomized estimate is cause for questioning the nonrandomized one. These claims are related but they are not identical. Widespread agreement that the first claim is true has perhaps led to uncritical acceptance that the second claim necessarily follows. Agreement on the first claim is well-justified based on the strong statistical theory buttressing the randomized experiment. But the second claim goes beyond that statistical theory in three senses. First, statistical theory prioritizes on effects from randomized
experiments on expectation. Any given comparison of random to nonrandom experiments may result in differences between them due to chance, with sampling error reducing the value of the gold standard in such cases. Second, statistical theory takes little account of the vagaries of practical implementation of experiments, from attrition and partial treatment implementation to the local embeddedness of any given experiment with particular times, persons, settings, operationalizations of treatment and outcomes. Given difficulty holding those vagaries perfectly constant over assignment methods, even in a highly controlled study like the present one, some discrepancies between the resulting estimates may be due to differences in implementation rather than to biased estimates. This is especially true given the near impossibility of knowing all these implementation problems, and the fact that they likely affect both randomized and nonrandomized experiments. Third, the randomized gold standard is problematic in a way particular to comparisons with RDD because they typically estimate two different parameters. We can force the RE to estimate ATEC, but RE-ATEC is a less precise estimate than RE-ATE given that it heavily weights only those observations close to the cutoff. Consequently, these three reasons might lead us to expect some differences between randomized and nonrandomized experiments even if both are yielding unbiased estimates of effect—differences in whether both estimates are in the same direction, of the same magnitude, both statistically significant or both not so, or significantly different from each other. What has received little serious attention in the literature is how much similarity in magnitude of effect is enough, or how many significant differences between estimates are acceptable.

For all these reasons, then, perhaps the discrepancies observed in the present study should not be cause to question the RDD estimates. Further, a more general implication may follow. The present paradigm is a special case of the within-study paradigm described by Cook et al. (2008)
in which results from randomized experiments are compared to those from nonrandomized
experiments, including but not limited to RDD. Less than a handful of studies in that paradigm
(e.g., Shadish et al., 2008; Pohl, Steiner, Eisermann, Soellner & Cook, 2009) have randomized
participants to be in a randomized or nonrandomized experiment. If imbalances in covariates
across assignment methods occur when participants are randomly assigned to assignment
method, as they did in the present study, is it not likely that even more covariate imbalance will
occur when they are not randomly assigned? Such imbalances might potentially cause even less
agreement than in the present study between results from randomized and nonrandomized
experiments. The implication is that past authors may have overstated the case when they
interpreted their results as suggesting that nonrandomized experiments may not well approximate
results from randomized experiments (e.g., Glazerman et al., 2003; Peikes, Moreno & Orzol,
2008).

**Generalizability**

Questions arise about the extent to which results from a laboratory analogue study like
the present one might generalize to other contexts. Studies that might show how the present
results might generalize have been limited. The present article slightly generalizes the Shadish et
al. (2008) methods by fielding an internet-administered study where the treatment
implementation problems described in the methods section occurred (more attrition, probably
less constant treatment effect, possibly some cheating) that did not do so in the more controlled
methods with other changes. They translated the study into German, used introductory
psychology or education students, changed the vocabulary training to a treatment designed to
improve the English of German speaking students, used some new covariates appropriate to the
German setting, changed the vocabulary outcome to one measuring English skills, and made a few minor technical changes. Also, the nature of the self-selection bias was different, with German students choosing English training when they were less proficient in English compared to Shadish et al.’s (2008) American students choosing vocabulary because they were more proficient at it than at math. Still, Pohl et al.’s results replicated Shadish et al. (2008). However, considerable ambiguity about generalizability remains. Shadish et al. (2008) described the conceptual issues likely to be involved in such generalizations, and they have not changed.

**Conclusion**

This discussion has focused mostly on the differences between RD and RE estimates in this study. This should not blind us to the fact that the two estimates were quite concordant on the whole. Estimates from the regression discontinuity design were always in the same direction, of the same magnitude, and consistent in rejecting the null hypothesis or not, as estimates from the randomized experiment. The estimates were not significantly different from each other when analyzed as usual, and when they were forced to estimate the same parameter, they were only clearly significantly different from each other in three of sixteen analyses. Randomized experiments are still preferable to regression discontinuity designs because they have more power and fewer assumptions. But researchers who need to use a regression discontinuity design can, given the present study and its predecessors, have reasonable confidence that they are getting an accurate estimate of the effects of treatments.
References


Figure 1. Graphical depiction of the assignment process and sample size. The number before the slash (/) is sample size initially assigned, and after the slash is sample size after attrition.
Figure 2. Scatterplot of the regression discontinuity data for the mathematics outcome with local linear regression lines.
Figure 3. Scatterplot of the randomized experiment for the mathematics outcome with local linear regression lines (the vocabulary group’s regression line is almost linear despite the non-parametric regression).
Figure 4. Scatterplot of the regression discontinuity data for the vocabulary outcome with local linear regression lines.
Figure 5. Scatterplot of the randomized experiment for the vocabulary outcome with local linear regression lines.
Table 1. Models for estimating treatment effects in the randomized experiment and regression discontinuity design.

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Linear regression (RE only)</td>
<td>( Y_i = \alpha + \beta_1 D_i + \beta_2 Z_i + X_i' \gamma + \epsilon_i )</td>
</tr>
<tr>
<td>(2) Linear regression</td>
<td>( Y_i = \alpha + \beta_1 D_i + \beta_2 Z_i + \beta_3 D_i Z_i + X_i' \gamma + \epsilon_i )</td>
</tr>
<tr>
<td>(3) Quadratic regression</td>
<td>( Y_i = \alpha + \beta_1 D_i + \beta_2 Z_i + \beta_3 D_i Z_i + \beta_4 D_i Z_i^2 + X_i' \gamma + \epsilon_i )</td>
</tr>
<tr>
<td>(4) Cubic regression</td>
<td>( Y_i = \alpha + \beta_1 D_i + \beta_2 Z_i + \beta_3 D_i Z_i + \beta_4 D_i Z_i^2 + \beta_5 D_i Z_i^3 + X_i' \gamma + \epsilon_i )</td>
</tr>
<tr>
<td>(5) Generalized Additive Model (GAM)</td>
<td>( Y_i = \tau D_i + f_o(Z_{0i}) + f_i(Z_{ui}) + X_i' \gamma + \epsilon_i )</td>
</tr>
<tr>
<td>(6) Local Linear Regression (at ( Z_i = 0 ))</td>
<td>( Y_i = \alpha + \tau D_i + \beta_1 Z_i + \beta_2 D_i Z_i + X_i' \gamma + \epsilon_i ) with triangular kernel weights ( K(Z_i/h) )</td>
</tr>
</tbody>
</table>

Notes. \( Y_i \) is the outcome for subjects \( i = 1, \ldots, N \); \( D_i \) is a treatment indicator with \( D_i = 1 \) for treatment and \( D_i = 0 \) for control (according to randomization in the randomized experiment or cutoff-based assignment in the regression discontinuity design) with \( \tau \) being the treatment effect of interest; Regression coefficient \( \alpha \) refers to the constant term, \( \beta \)'s indicate the effect of the polynomial of the centered assignment variable \( Z_i \) and the interaction with treatment \( D_i \); Vector \( \gamma \) represents the coefficient vector for the column vector of all measured predictors \( X_i \) (see Table 4). For the second set of exploratory analyses, \( X_i \) also includes the treatment \( \times \) math-pretest interaction (with the math pretest centered at the estimated pretest value at the reading cutoff of the RDD data). For all models, except for the local linear regression, error terms \( \epsilon_i \), for \( i = 1, \ldots, N \), are assumed to be independent and normally distributed.

In estimating parametric models (2)-(4) we also accounted for specification error due to the discreteness of the assignment variable (Lee & Card, 2008). Since correcting for the clustering effect resulted in reduced standard errors, we report the larger conventional regression standard errors (Angrist & Pischke, 2009).
In the generalized additive model, \( f_0(Z_{0i}) \) and \( f_1(Z_{1i}) \) are smooth functions of the assignment variable for the control and treatment group, respectively, with \( Z_{0i} = (1 - D_i)Z_i \) and \( Z_{1i} = D_iZ_i \). They are estimated via thin plate regression splines using the \textit{mcgv}-package in R (R Development Core Team, 2009; Wood, 2006).

The local linear regression (Imbens & Lemieux, 2008) uses a triangular kernel \( K(Z_i/h) \) centered at the cutoff with an optimal selection of bandwidth \( h \) according to Imbens & Kalyanaraman (2010). The optimal bandwidth was determined for the RDD data and held constant for the RE data.
Table 2. Effect of Mathematics Training on Mathematics Outcome in the Randomized Experiment (RE) and the Regression Discontinuity Design (RDD).

<table>
<thead>
<tr>
<th>Mathematics</th>
<th>RE Mean (SE)</th>
<th>RDD Mean (SE)</th>
<th>Difference Between RDD and RE Difference (SE)</th>
<th>Effect Size</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent Analyses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) RE-ATE vs (2) RDD ATEC</td>
<td>2.53 (0.60)***</td>
<td>1.87 (0.59)***</td>
<td>-0.66 (0.85)</td>
<td>-0.15</td>
<td>-0.78</td>
<td>0.43</td>
</tr>
<tr>
<td><strong>Exploratory Analyses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Linear regression</td>
<td>2.64 (0.60)***</td>
<td>1.87 (0.60)***</td>
<td>-0.64 (0.85)</td>
<td>-0.17</td>
<td>-0.91</td>
<td>0.36</td>
</tr>
<tr>
<td>(3) Quadratic regression</td>
<td>2.70 (0.73)***</td>
<td>1.88 (0.83)*</td>
<td>-0.83 (1.10)</td>
<td>-0.18</td>
<td>-0.75</td>
<td>0.45</td>
</tr>
<tr>
<td>(4) Cubic regression</td>
<td>2.96 (0.75)***</td>
<td>2.64 (1.13)*</td>
<td>-0.32 (1.36)</td>
<td>-0.07</td>
<td>-0.23</td>
<td>0.81</td>
</tr>
<tr>
<td>(5) Generalized additive model (GAM)</td>
<td>3.24 (0.75)***</td>
<td>1.87 (0.60)***</td>
<td>-1.37 (0.96)</td>
<td>-0.31</td>
<td>-1.43</td>
<td>0.15</td>
</tr>
<tr>
<td>(6) Local Linear Regression</td>
<td>2.69 (1.36)*</td>
<td>1.98 (0.86)*</td>
<td>-0.71 (1.61)</td>
<td>-0.16</td>
<td>-0.44</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Exploratory Analyses with Math Pretest Interaction Covariate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Linear regression</td>
<td>2.57 (0.61)***</td>
<td>1.94 (0.59)***</td>
<td>-0.64 (0.85)</td>
<td>-0.14</td>
<td>-0.74</td>
<td>0.46</td>
</tr>
<tr>
<td>(3) Quadratic regression</td>
<td>2.61 (0.75)***</td>
<td>1.94 (0.82)*</td>
<td>-0.68 (1.11)</td>
<td>-0.15</td>
<td>-0.61</td>
<td>0.54</td>
</tr>
<tr>
<td>(4) Cubic regression</td>
<td>2.88 (0.78)***</td>
<td>2.81 (1.13)**</td>
<td>-0.07 (1.37)</td>
<td>-0.02</td>
<td>-0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>(5) Generalized additive model (GAM)</td>
<td>3.19 (0.77)***</td>
<td>1.94 (0.59)**</td>
<td>-1.25 (0.97)</td>
<td>-0.28</td>
<td>-1.19</td>
<td>0.20</td>
</tr>
<tr>
<td>(6) Local Linear Regression</td>
<td>2.63 (1.35)</td>
<td>2.06 (0.86)*</td>
<td>-0.57 (1.61)</td>
<td>-0.13</td>
<td>-0.35</td>
<td>0.72</td>
</tr>
</tbody>
</table>

* \( p < .05; \) ** \( p < .01; \) *** \( p < .001 \)

Notes: Effect sizes are calculated as the RDD-RE difference divided by the pooled standard deviation of the randomized treatment and control group. Standard errors of the local linear regression estimates are bootstrapped from 1000 samples. \( t = (\hat{\tau}_{RDD} - \hat{\tau}_{RE})/SE_{RDD-RE} \) indicates the \( t \)-statistic of the RDD-RE difference with standard error (SE) \( SE_{RDD-RE} = \sqrt{SE_{RDD}^2 + SE_{RE}^2} \). \( p \) is the corresponding level of empirical significance using the standard normal distribution as an approximation (sample sizes are 380 and 189 for the RDD and RE, respectively).
Table 3. Effect of Vocabulary Training on Vocabulary Outcome in the Randomized Experiment (RE) and the Regression Discontinuity Design (RDD).

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>RE Mean (SE)</th>
<th>RDD Mean (SE)</th>
<th>Difference (SE)</th>
<th>Effect Size</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent Analyses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) RE-ATE vs (3) RDD ATEC</td>
<td>4.46 (0.49)**</td>
<td>5.95 (0.78)**</td>
<td>1.49 (0.92)</td>
<td>0.41</td>
<td>1.62</td>
<td>0.10</td>
</tr>
<tr>
<td>Exploratory Analyses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Linear regression</td>
<td>4.35 (0.49)**</td>
<td>5.10 (0.57)**</td>
<td>0.75 (0.75)</td>
<td>0.21</td>
<td>1.00</td>
<td>0.32</td>
</tr>
<tr>
<td>(3) Quadratic regression</td>
<td>3.92 (0.58)**</td>
<td>5.95 (0.78)**</td>
<td>2.03 (0.97)*</td>
<td>0.56</td>
<td>2.09</td>
<td>0.04</td>
</tr>
<tr>
<td>(4) Cubic regression</td>
<td>3.64 (0.60)**</td>
<td>6.16 (1.07)**</td>
<td>2.51 (1.23)*</td>
<td>0.69</td>
<td>2.05</td>
<td>0.04</td>
</tr>
<tr>
<td>(5) Generalized additive Model (GAM)</td>
<td>2.74 (0.70)**</td>
<td>5.81 (0.72)**</td>
<td>3.08 (1.01)*</td>
<td>0.84</td>
<td>3.06</td>
<td>0.01</td>
</tr>
<tr>
<td>(6) Local Linear Regression</td>
<td>3.89 (1.53)**</td>
<td>6.19 (0.96)**</td>
<td>2.29 (1.77)</td>
<td>0.63</td>
<td>1.29</td>
<td>0.20</td>
</tr>
<tr>
<td>Exploratory Analyses with Math Pretest Interaction Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Linear regression</td>
<td>4.55 (0.49)**</td>
<td>5.05 (0.57)**</td>
<td>0.50 (0.75)</td>
<td>0.14</td>
<td>0.67</td>
<td>0.51</td>
</tr>
<tr>
<td>(3) Quadratic regression</td>
<td>4.15 (0.59)**</td>
<td>5.91 (0.77)**</td>
<td>1.76 (0.98)</td>
<td>0.48</td>
<td>1.80</td>
<td>0.07</td>
</tr>
<tr>
<td>(4) Cubic regression</td>
<td>3.89 (0.62)**</td>
<td>6.04 (1.07)**</td>
<td>2.15 (1.24)</td>
<td>0.59</td>
<td>1.74</td>
<td>0.08</td>
</tr>
<tr>
<td>(5) Generalized additive model (GAM)</td>
<td>3.92 (0.82)**</td>
<td>5.76 (0.72)**</td>
<td>1.84 (1.09)</td>
<td>0.50</td>
<td>1.68</td>
<td>0.09</td>
</tr>
<tr>
<td>(6) Local Linear Regression</td>
<td>4.11 (1.61)**</td>
<td>6.15 (0.96)**</td>
<td>2.04 (1.85)</td>
<td>0.56</td>
<td>1.10</td>
<td>0.27</td>
</tr>
</tbody>
</table>

* p < .05; ** p < .01; *** p < .001

Notes: Effect sizes are calculated as the RDD-RE difference divided by the pooled standard deviation of the randomized treatment and control group. Standard errors of the local linear regression estimates are bootstrapped from 1000 samples. $t = (\hat{\tau}_{RDD} - \hat{\tau}_{RE}) / SE_{RDD-RE}$ indicates the $t$-statistic of the RDD-RE difference with standard error (SE) $SE_{RDD-RE} = \sqrt{SE_{RDD}^2 + SE_{RE}^2}$; $p$ is the corresponding level of empirical significance using the standard normal distribution as an approximation (sample sizes are 380 and 189 for the RDD and RE, respectively).
Table 4. Pretest Differences Between Randomized Experiment and Regression Discontinuity Design.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Overall mean difference</th>
<th>Mean difference at cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d$</td>
<td>$t$</td>
</tr>
<tr>
<td>Vocabulary Pretest$^1$</td>
<td>-0.07</td>
<td>-0.78</td>
</tr>
<tr>
<td>Mathematics Pretest</td>
<td>0.26</td>
<td>2.92**</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.03</td>
<td>-0.31</td>
</tr>
<tr>
<td>Age 18</td>
<td>0.02</td>
<td>0.24</td>
</tr>
<tr>
<td>Age 19</td>
<td>-0.01</td>
<td>-0.10</td>
</tr>
<tr>
<td>Age 20</td>
<td>-0.09</td>
<td>-1.02</td>
</tr>
<tr>
<td>Age 21</td>
<td>0.08</td>
<td>0.87</td>
</tr>
<tr>
<td>Ethnicity Asian</td>
<td>0.08</td>
<td>0.94</td>
</tr>
<tr>
<td>Ethnicity African American</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>Ethnicity Hispanic</td>
<td>-0.13</td>
<td>-1.49</td>
</tr>
<tr>
<td>Ethnicity Caucasian</td>
<td>-0.01</td>
<td>-0.10</td>
</tr>
<tr>
<td>Ethnicity Other</td>
<td>0.10</td>
<td>1.16</td>
</tr>
<tr>
<td>Have Taken High School Algebra</td>
<td>0.04</td>
<td>0.47</td>
</tr>
<tr>
<td>Have Taken High School Calculus</td>
<td>0.02</td>
<td>0.20</td>
</tr>
<tr>
<td>Have Taken College Algebra</td>
<td>0.07</td>
<td>0.80</td>
</tr>
<tr>
<td>Have Taken College Calculus</td>
<td>0.01</td>
<td>0.12</td>
</tr>
<tr>
<td>Like Engineering</td>
<td>-0.01</td>
<td>-0.15</td>
</tr>
<tr>
<td>Like Math</td>
<td>-0.14</td>
<td>-1.55</td>
</tr>
<tr>
<td>Will Take Extra Math</td>
<td>-0.12</td>
<td>-1.31</td>
</tr>
<tr>
<td>Will Take Extra Engineering</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>Math Intensive Major</td>
<td>0.05</td>
<td>0.59</td>
</tr>
<tr>
<td>Literature Intensive Major</td>
<td>-0.10</td>
<td>-1.07</td>
</tr>
<tr>
<td>No Declared Major</td>
<td>0.06</td>
<td>0.71</td>
</tr>
<tr>
<td>Like Math Best</td>
<td>-0.10</td>
<td>-1.09</td>
</tr>
<tr>
<td>Like Math and Literature Equally</td>
<td>0.09</td>
<td>1.01</td>
</tr>
<tr>
<td>Like Literature Best</td>
<td>0.02</td>
<td>0.19</td>
</tr>
</tbody>
</table>

** p < .01

1 Test of mean difference at cutoff was not done because this variable was the assignment variable in the regression discontinuity design and so must display a difference.

Note. $d = (\bar{X}_{RDD} - \bar{X}_{RE}) / \sqrt{(SD^2_{RDD} + SD^2_{RE})/2}$ is the standardized mean difference where SD is the standard deviation of covariate $X$ in the regression discontinuity design or randomized experiment. $t = (\bar{X}_{RDD} - \bar{X}_{RE}) / \sqrt{SE^2_{RDD} + SE^2_{RE}}$ indicates the two independent samples $t$-statistic of the RDD-RE difference with Welch’s corrected degrees of freedom ($df$). For the mean difference at the cutoff calculations included only observations with an assignment score close to the cutoff, that is, 3 points below or above the cutoff.