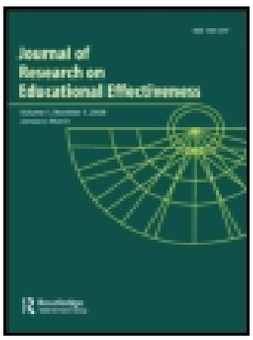

Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Using Data from Randomized Trials to Assess the Likely Generalizability of Educational Treatment-Effect Estimates from Regression Discontinuity Designs

Howard Bloom, Andrew Bell & Kayla Reiman

To cite this article: Howard Bloom, Andrew Bell & Kayla Reiman (2020): Using Data from Randomized Trials to Assess the Likely Generalizability of Educational Treatment-Effect Estimates from Regression Discontinuity Designs, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2019.1634169

To link to this article: https://doi.org/10.1080/19345747.2019.1634169

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2019.1634169

METHODOLOGICAL STUDIES

Using Data from Randomized Trials to Assess the Likely Generalizability of Educational Treatment-Effect Estimates from Regression Discontinuity Designs

Howard Bloom[a], Andrew Bell[a], and Kayla Reiman[a]

ABSTRACT

This article assesses the likely generalizability of educational treatment-effect estimates from regression discontinuity designs (RDDs) when treatment assignment is based on academic pretest scores. Our assessment uses data on outcome and pretest measures from six educational experiments, ranging from preschool through high school, to estimate RDD generalization bias. We then compare those estimates (reported as standardized effect sizes) with the What Works Clearinghouse (WWC) standard for acceptable bias size ($\leq 0.05\sigma$) for two target populations, one spanning a half–standard deviation pretest-score range and another spanning a full–standard deviation pretest-score range. Our results meet this standard for all 18 study/outcome/pretest scenarios examined given the narrower target population, and for 15 scenarios given the broader target population. Fortunately, two of the three exceptions represent pronounced "ceiling effects" that can be identified empirically, making it possible to avoid unwarranted RDD generalizations, and the third exception is very close to the WWC standard.

KEYWORDS

regression discontinuity
generalizability
validation test

Introduction

In recent years, the regression discontinuity design (RDD) has gained widespread recognition as a quasi-experimental method that can produce internally valid estimates of causal effects of a treatment, a program or an intervention—hereafter referred to as treatment effects. Consequently, RDDs have been used to estimate causal effects in a variety of fields (Lee & Lemieux, 2010 provide a list of more than 75 RDD studies of education, labor markets, political economy, health care, and criminal justice). Research on the statistical properties of RDDs has theoretically justified and empirically verified their considerable potential for internal validity (Chaplin et al., 2018; Hahn, Todd, & Van Der Klaauw, 2001; Imbens & Lemieux, 2008; Lee, 2008; Lee & Lemieux, 2010).

Although various, sometimes complex, estimation strategies have been used to implement RDDs (see Lee & Lemieux, 2010 for a review), the logic of RDD treatment-effect identification is straightforward and intuitively appealing. First, an RDD applies to

CONTACT Howard Bloom howard.bbloom@mdrc.org MDRC, 200 Vesey Street, 23rd Floor, New York, NY 10281-2103, USA.
[a]MDRC, New York, New York, USA

Supplemental data for this article can be accessed on the publisher's website.

situations in which subjects are assigned to a treatment based on whether their value for a numeric rating exceeds (or falls below) a specified threshold or "cut-point." This assignment method can produce an abrupt shift or discontinuity at the cut-point in the probability of assignment to treatment, which in turn can produce a corresponding discontinuity in mean outcomes. Under a set of assumptions which are often plausible, the discontinuity in the mean outcome at the RDD cut-point equals the local mean causal effect of treatment assignment.

The assumptions needed to establish internal validity for this identification strategy are: (1) independence between determination of the RDD cut-point and determination of subjects' ratings, (2) local continuity of the functional relationships between mean outcomes and ratings on both sides of the RDD cut-point, and (3) local discontinuity in the probability of treatment assignment at the cut-point (Hahn et al., 2001). Furthermore, random measurement error (noise) in observed ratings can make an RDD the approximate statistical equivalent of a local randomized controlled trial (RCT) at the RDD cut-point (Lee & Lemieux, 2010). Hence, there are good reasons to expect a well-implemented RDD to have strong internal validity.

While it is true that without further assumptions, RDD estimates of treatment effects are identifiable only at the RDD cut-point,[1] this does not necessarily mean that such estimates only apply to a narrow, homogeneous, and potentially nonrelevant subpopulation. For this limitation to hold with force, treatment effects must vary widely across the target population of interest for an RDD study *and* they must co-vary strongly with observed ratings. If treatment effects do not vary widely or if they do not co-vary strongly with observed ratings, the conditional distribution of individual treatment effects at an RDD cut-point can be similar to the unconditional distribution of individual treatment effects across the range of ratings for the target population. Unfortunately, little is known about the extent to which treatment effects vary across individuals and even less is known about predictors of this variation.[2] Consequently, there is little existing guidance for deciding when and how far to generalize RDD findings.

To help inform such judgements, this article explores how treatment effects for participants in six major educational interventions, which run the developmental gamut from preschool to high school, co-vary with participants' academic pretest score (a measure that is often a basis for treatment assignment in education). In what follows, we: (1) reflect on the factors that could challenge the generalizability of RDD findings, (2) review recent approaches for assessing or enhancing this generalizability, (3) describe our empirical approach for exploring this generalizability, (4) present key findings produced by this approach, and (5) consider the implications, strengths, and limitations of those findings.

The RDD Generalizability Challenge

Variation in treatment effects is a growing area of research, and recent work has focused on how to measure this variation—across subgroups defined by individual background

[1]Some early RDD analyses (see Cook & Campbell, 1979 for a discussion) used what are now considered to be strong assumptions about the functional form of the relationship between mean outcomes and ratings to generalize treatment-effect estimates beyond an RDD cut-point.

[2]For example, Tipton, Yeager, Iachan, and Schneider (2019) note the frequent inability to replicate experimentally estimated treatment-effect differences for demographic subgroups.

Definitions:

ρ indicates the reliability of the observed rating. This is the proportion of total variation in a measure that reflects true variation in whatever is being measured systematically.

β represents the rate at which the conditional mean treatment effect, τ(R*), changes per unit change in the true rating, R*.

ρβ is defined as the strength of the linear relationship between observed rating (R) and treatment effect (τ). This is the coefficient of a bivariate linear regression of τ on R.

Figure 1. The relationship between treatment effects and RDD ratings.

characteristics (Bloom & Michalopoulos, 2011; Rothwell, 2005), across research studies using methods of meta-analysis (Cooper & Hedges, 1994), and across research sites using multilevel models (Bloom et al., 2017; Weiss et al., 2017). This work illustrates that studying variation in treatment effects is challenging. For example, to accurately estimate variation in treatment effects across sites requires a substantial number of sites with large samples (Bloom & Spybrook, 2017). Furthermore, it is even more difficult to study variation in treatment effects across individuals because without strong assumptions it is not possible to identify individual-level treatment effects (Bloom et al., 2017; Raudenbush & Bloom, 2015). Consequently, although there is now some information about how much treatment effects vary across sites (Weiss et al., 2017) there is little information about this variation across individuals.

However, when assessing the generalizability of RDD findings, one does not need to consider *all* variation in treatment effects that might exist. It is only necessary to consider treatment-effect variation that *co-varies* with the RDD rating. Specifically, the concern for generalizing RDD findings is that mean treatment effects might differ substantially across rating values. However, this concern is only justified if the mix of subjects differs substantially across rating values in terms of factors that predict treatment effects. Figure 1 illustrates the relationships that must exist for this to be the case.

Note first the relationship between observed and true values of an RDD rating (R and R^*, respectively). A subject's true rating is the actual value of his ability, merit, disadvantage, need, or whatever characteristic is the basis for treatment assignment.[3] However, this true rating is typically unobservable and can only be approximated by an observed rating, which contains random measurement error or noise. There are many sources of such noise. For example, ratings based on test scores contain noise from inadvertent mistakes, memory lapses, and ambiguous question-wording; and ratings based on classroom observations or grader assessments contain noise due to variation in observer or grader judgment. Furthermore, the attitudes, behavior, or performance of subjects on a measure can be affected by random events in their lives such as illness, job loss, or stressful changes in personal circumstances. Even ratings that reflect seemingly perfectly

[3]By true rating we mean whatever an RDD rating measures systematically, regardless of what it is intended to measure. Hence, we focus on the reliability of an RDD rating, not its validity.

knowable facts, like students' birthdays, can contain random measurement error due to coding mistakes.

As noted above, noise in an RDD rating improves the degree to which an RDD can approximate a local RCT at the RDD cut-point, which enhances the internal validity of RDD findings.[4] This reflects the fact that, for subjects with the same true rating value, it is a matter of chance whether their observed rating falls above or below the RDD cut-point. Less widely known, but perhaps equally important, is Lee and Lemieux's (2010) insight that noise in an observed rating is critical for the external validity or generalizability of RDD findings. As they explain, random error in observed ratings produces heterogeneity of subjects' true rating for each observed value of the rating, including its cut-point value. For example, because of noise, students with varying true past achievement levels (and true treatment effects) might, by chance, receive the same score on a pretest (rating) that assigns them to mandatory summer school if they fail the test. Other things being equal then, increased noise in a measure implies greater heterogeneity of the cut-point subpopulation, which in turn implies greater generalizability of RDD findings. In the extreme, if observed ratings consist solely of noise, their values are random numbers and individuals assigned to treatment based on those numbers are assigned randomly, as in an RCT (Bloom, 2012; Lee & Lemieux, 2010).

However, noise in observed ratings—the focus of Lee and Lemieux's insight—is only one of two factors that can influence the generalizability of RDD findings. As Figure 1 illustrates, the relationship between observed ratings and treatment effects is the *product* of two intervening statistical relationships, or "links in a predictive chain," each of which can weaken the overall relationship between mean treatment effects and observed ratings and thereby improve RDD generalizability. We refer to this relationship as a predictive chain, with nondirectional lines, instead of a causal chain, with directional arrows, because it is not necessary to identify causality for the present discussion of predictive relationships.

The first link in the predictive chain is between observed and true ratings (R and R^*). It is convenient to summarize the strength of this link as the reliability (ρ) of the observed rating, where reliability (e.g., Brennan, 2001; Nunnally, 1967) is the proportion of total variation in a measure that reflects true variation in whatever is being measured systematically. This parameter can range from a value of one if there is no noise in the observed rating and all variation in it represents true variation, to a value of zero if there is no information in the observed rating and all variation in it represents noise.

The second link in the predictive chain represented by Figure 1 is the relationship between true ratings and treatment effects (R^* and τ, respectively). It is convenient to represent the strength of this link as the slope of a linear regression of τ (the dependent variable) on R^* (the predictor). This slope represents the rate at which the conditional mean treatment effect, $\tau(R^*)$, changes per unit change in the true rating, R^*. As described later, we also consider a quadratic relationship between intervention effects and pretest scores.

In education and human development, RDD ratings are often measures of past outcomes, which are used to assess the need for or ability to benefit from a specific

[4]This approximation improves as the interval around the RDD cut-point decreases.

intervention. However, there is little consistent evidence about how or how well the constructs which underlie these measures predict treatment effects. For example, Cunha and Heckman (2007) hypothesize that "skills beget skills," which implies that students' ability to experience positive effects from a developmental initiative increases with the strength of the backgrounds they bring to it. In contrast, other researchers (e.g., Ramey & Ramey, 1998) hypothesize a "compensatory" pattern of impacts for initiatives that focus on countering educational disadvantages. In addition, for binary outcomes like achieving proficiency on a mandatory test, there has been extensive debate about "bubble kids" who are close to being able to surmount a required threshold and thus might benefit most from assistance (Booher-Jennings, 2005). One could also imagine an analogous "sweet spot" theory for continuous outcomes, in which the students who benefit most from an intervention have strong enough educational backgrounds to participate effectively but sufficient room for improvement to exhibit demonstrable gains. However, because it has been difficult to replicate treatment-effect differences among subgroups of individuals (e.g., Tipton et al., 2019), it is unclear how or how strongly educational treatment effects co-vary with students' background characteristics.

Together, the relationship between observed and true ratings (R and R^*) and the relationship between true ratings and treatment effects (R^* and τ) determine the strength of the relationship between observed ratings and treatment effects (R and τ). The strength of this relationship ($\rho\beta$)[5] is the rate at which the conditional mean treatment effect changes per unit change in the observed rating. Thus, weakness in either of the intervening links (ρ or β) weakens the overall relational chain. Furthermore, weaknesses in these links are compounding and thus accumulate rapidly, which can increase RDD generalizability.

One further issue to consider when exploring the generalizability of an RDD treatment-effect estimate is the appropriate target population for that estimate. For example, most interventions are designed to address a specific problem. Hence, they are often intended for a subpopulation that is especially at risk of having the problem or is currently experiencing it. In these cases, the ability of an RDD to generalize findings to that subpopulation is more relevant for policy and practice than is its ability to generalize to a more general population. Consider, for example, the Double-Dose Algebra initiative in Chicago (Nomi & Raudenbush, 2016) to which high school students with math test scores below a specified level were assigned. For interventions like this, it is most relevant to generalize evaluation findings to a *subpopulation* of struggling students, which should be taken into account when assessing the practical generalizability of RDD findings.

Recent Approaches to RDD Generalizability

Three recent articles have made important advances in the assessment and enhancement of RDD generalizability: Wing and Cook (2013), Angrist and Rokkanen (2015), and

[5]This result reflects the well-known phenomenon of attenuation bias or errors-in-variables bias in an estimated regression coefficient due to random error in the independent variable for that coefficient (e.g., Angrist & Pischke, 2015, pp. 240–241; Wooldridge, 2009, p. 320).

Dong and Lewbel (2015).[6] Wing and Cook (2013) propose an approach which uses information on an outcome that is measured before and after treatment assignment by the same instrument and on the same scale (a generic pretest and post-test) to estimate treatment effects at and beyond an RDD cut-point. Their approach applies when pre-tests and post-tests measure the same thing the same way (Medicaid expenditures from administrative records before and after treatment assignment in their empirical example) and the RDD rating measures something different (sample members' age at treatment assignment in their example).

Intuitively, the logic of the Wing and Cook (2013) approach is as follows. First, for RDD comparison group members, they compare the statistical relationship between pre-tests and ratings with that between post-tests and ratings and model the difference, preferably with a simple intercept shift. Second, for RDD treatment group members, they use the estimated model from step 1 to impute counterfactual untreated post-test values from observed pretest values. Third, they estimate the treatment effect for RDD treatment group members as the difference between their actual post-test values and their imputed counterfactual post-test values. The key assumptions of this approach are that: (1) it is possible to model the difference between the pretest/rating relationship and the post-test/rating relationship for comparison group members within a specified distance from the RDD cut-point and (2) this modeled difference also holds for treatment group members within that specified distance from the cut-point.

The authors test their approach by constructing the equivalent of conventional RDD samples and pretest-enhanced RDD samples using experimental data from the Cash and Counseling Demonstration Experiment for Medicaid recipients (Brown & Dale, 2007). They construct a given RDD from the experimental data by: (1) choosing an illustrative cut-point for the emulated RDD rating (sample members' age at experimental treatment assignment), (2) defining RDD comparison group members as experimental control group members whose age at treatment assignment was below this cutoff, and (3) defining RDD treatment group members as experimental treatment group members whose age at treatment assignment was at or above the cut-point. This was done for three age cutoffs separately for each of the three states in the study they used.

Then, for each of these nine analysis samples, the authors use an intercept shift to model the difference between the pretest/rating relationship and the post-test/rating relationship. By comparing treatment-effect estimates and their standard errors for the two types of RDDs with their experimental counterparts, the authors found that the pre-test-enhanced RDD outperformed the conventional RDD in terms of bias and precision and produced results that were broadly comparable to their experimental benchmarks—both at and beyond the emulated treatment-assignment threshold.

Angrist and Rokkanen (2015) propose a method for estimating mean treatment effects away from an RDD cut-point. Their approach builds on the fundamental fact

[6]Chaplin et al. (2018) examine the internal validity of RDD treatment-effect estimates in practice through a meta-analysis of 15 within-study comparisons of RDD findings to their RCT counterparts. In doing so, the authors estimate the mean and variation of RDD bias across studies. They use the mean RDD bias as a summary measure of the internal validity of the method for the range of situations examined. They use variation in assessed RDD bias as a measure of the external validity of the estimated mean bias for assessing the likely internal validity of a specific RDD. Their focus on the external validity of a *bias assessment* of the RDD method differs from the present focus on the likely generalizability of RDD *treatment-effect estimates*.

that the only possible source of selection bias in treatment-assignment for a properly implemented RDD is the relationship between sample members' ratings and their potential outcome values. Thus, if one can eliminate this relationship by controlling for sample members' background characteristics that predict ratings and potential outcome values, one can produce *conditionally ignorable* treatment assignment. This in turn makes it possible to use conventional regression or propensity-score methods to estimate mean treatment effects for any rating value or range of rating values.

As Angrist and Rokkanen (2015) demonstrate, it is possible to test this conditional ignorability assumption (CIA) empirically by regressing (separately for treatment group members and comparison group members) the outcome measure of interest on the RDD rating plus a set of covariates. If the estimated regression coefficient for the rating is small and not statistically significant, this suggests that residual values of the rating (controlling for the covariates) are uncorrelated with residual values of potential outcomes (controlling for the covariates), which implies that treatment assignment is conditionally ignorable. If this is the case—and if for any given set of covariate values, there are sample members assigned to treatment and others assigned to comparison status—then straightforward nonexperimental estimators can be used to estimate mean treatment effects.[7]

Angrist and Rokkanen (2015) used their approach to estimate effects of assignment to two of Boston's elite exam schools on students' tenth-grade state test scores in math and English language arts (ELA). They did so for students who applied to enter the schools in seventh grade and for students who applied to enter the schools in ninth grade. Covariates used to approximate conditional ignorability were pretest scores plus standard demographic, social, and economic indicators.[8] Because pretest scores for seventh-grade applicants were more limited than those for ninth-grade applicants, and thus empirical results for conditional ignorability were more encouraging for tenth-grade applicants, the authors only report impact findings for ninth-grade applicants. These results suggest that mean treatment effects away from the RDD cut-point were similar to those at the cut-point for both schools studied.

Dong and Lewbel (2015) propose a method for using standard RDD methods to estimate the rate at which the mean treatment effect at an RDD cut-point changes per unit change in the RDD rating. The authors call this parameter a treatment effect derivative (TED) and use it to address two evaluation questions: (1) how well do RDD treatment-effect estimates generalize to subjects with ratings near the RDD cut-point? and (2) how would marginally changing the treatment assignment threshold change the mean treatment effect at the threshold?

For functional relationships between mean outcomes and ratings that are both continuous and continuously differentiable immediately below and above an RDD cut-point, Dong and Lewbel (2015) prove that a TED equals the *difference* between the first derivatives of those relationships at the cut-point.[9] The authors then illustrate how to estimate a

[7]The authors also note that a second empirical test of the validity of such nonexperimental estimators can be constructed by comparing a properly weighted nonexperimental estimate of the mean treatment effect at the RDD cut-point with its RDD counterpart. This weighting must match the distribution of covariate values at the RDD cut-point.
[8]These covariates were not part of the rating variable used to assign students to the elite exam schools.
[9]The authors demonstrate for continuous and continuously differentiable outcome/rating functions immediately below and above an RDD cut-point, with a mean outcome under the treated condition of $E(Y_1)$, a mean outcome under the

TED for a local-linear RDD regression with separate intercepts and slopes below and above the RDD cut-point and use this result to address their first research question.[10]

Dong and Lewbel (2015) extend this finding to situations with "local policy invariance" (where an intervention, its selection process, and its environment do not change in response to a marginal change in the treatment-assignment threshold) and demonstrate that in such cases, TED equals the rate of change in the mean treatment effect at the treatment-assignment threshold for an incremental change in that threshold. The authors call this parameter a marginal threshold treatment effect or MTTE and use it to address their second research question.

To illustrate their approach, Dong and Lewbel (2015) reanalyze RDD data from Goodman (2008) on the effect of Massachusetts' Adams Scholarship offers on student enrollment in a Massachusetts state college, which is tuition free for scholarship winners. Adams Scholarships are offered to high school graduates who exceed a specified threshold on the Massachusetts Comprehensive Assessment System (MCAS). The estimated treatment effect derivative for Adams Scholarships indicates that their effect on student enrollment drops precipitously for MCAS scores above the treatment assignment threshold (perhaps, as suggested by the authors, because students with these higher scores have more college options). In addition, the authors provide a plausible argument that Adams Scholarships are policy invariant, and thus TED equals MTTE for them. This implies that marginally raising the treatment assignment threshold would reduce the scholarship effect and marginally lowering the threshold would increase the scholarship effect, which illustrates the policy relevance of such findings.

The Present Approach

Unlike the preceding approaches which use RDD data to explore the *actual* generalizability of treatment-effect estimates produced from that data, we use RCT data to explore the *likely* generalizability of RDDs in a specific field: education research. Our approach is premised on the fact that ratings for many education RDDs are based in whole or in part on an academic pretest, like the math test scores used to assign students to Double-Dose Algebra in Chicago. Hence, the issue of RDD generalizability in educational research often translates into the question: *How and how strongly do treatment effects co-vary with students' pretest scores?* To address this question, our empirical strategy uses existing data from six educational RCTs.

Data

The six RCTs that we used were chosen from the 16 RCTs used by Weiss et al. (2017) to quantify cross-site impact variation for education and human development

untreated condition of $E(Y_0)$, an observed rating of R, and a mean treatment effect of τ (all at the RDD cut-point) that $\frac{d\tau}{dR} = \frac{dE(Y_1)}{dR} - \frac{dE(Y_0)}{dR}$.

[10]Consider the following local linear regression for a given bandwidth around an RDD cut-point with a uniform kernel. $Y_i = \alpha + \beta R_i + \gamma T_i + \delta T_i R_i + e_i$ where: R_i is the rating for subject i, T_i equals one if subject i was assigned to treatment and zero otherwise, and e_i is a random error that is distributed independently and identically across subjects with a mean of zero. In this case $\frac{d\tau}{dR} = \frac{dE(Y_1)}{dR} - \frac{dE(Y_0)}{dR} = (\beta + \delta) - \beta = \delta$.

Table 1. Study/outcome/pretest scenarios.

Study & outcome	Pretest	Sample size	Full sample impact estimate	*p* value for impact estimate
Head Start Impact Study				
Receptive vocab. score (z)	Letter/word score (*z*)	3,530	0.161***	0.000
Letter/word score (*z*)	Letter/word score (*z*)	3,540	0.218***	0.000
Oral comp. score (*z*)	Letter/word score (*z*)	3,480	0.017	0.489
After School Math				
Math score (*z*)	Math score (*z*)	1,970	0.075**	0.005
After School Reading				
Reading score (*z*)	Reading score (*z*)	1,890	−0.013	0.578
Enhanced Reading Opportunities				
% of required courses passed	Comp. score (*z*)	5,230	0.581*	0.015
Reading vocab. score (*z*)	Comp. score (*z*)	4,580	0.022	0.238
Reading comp. score (*z*)	Comp. score (*z*)	4,580	0.067***	0.000
Early College High Schools				
9th grade on track (%)	Math score (*z*)	3,390	6.300***	0.000
9th grade on track (%)	Reading score (*z*)	3,710	6.470***	0.000
Graduated (%)	Math score (*z*)	2,540	4.654**	0.001
Graduated (%)	Reading score (*z*)	2,720	4.031**	0.004
Small Schools of Choice				
9th grade credits accumulated (#)	Math score (*z*)	15,280	0.780***	0.000
9th grade credits accumulated (#)	Reading score (*z*)	14,730	0.776***	0.000
9th grade on track (%)	Math score (*z*)	15,870	8.248***	0.000
9th grade on track (%)	Reading score (*z*)	15,290	8.423***	0.000
Graduated (%)	Math score (*z*)	12,950	4.979***	0.000
Graduated (%)	Reading score (*z*)	12,460	5.198***	0.000

Source: Findings in this table are based on our computations using data from the studies represented. The full-sample mean treatment effect for each study/outcome/pretest scenario was estimated using Equation 1 from the present article with data for the scenario's parametric analysis sample.
Note. A two-tailed test was used to assess the statistical significance of each full-sample treatment-effect estimate with significance levels indicated by *p* values and by ***≤ 0.001, **≤ 0.01 and *≤ 0.05.

interventions. We selected these RCTs because we had worked with their data previously. They have at least one academic outcome and pretest; they have a large sample; they have acceptable attrition; and they have excellent treatment/control group pretest balance.[11]

The Appendix at the end of this article presents a synopsis of each RCT and Table 1 summarizes their key features. Note first that the interventions studied span a developmental range from preschool through high school. Preschool is represented by the national Head Start Impact Study (HSIS), which was conducted in a national sample of over 300 Head Start centers. Elementary school is represented by the After-School Math (ASM) demonstration and the After-School Reading (ASR) demonstration, which were each conducted in 25 after-school programs from across the United States. Last, high school is represented by the Enhanced Reading Opportunities (ERO) demonstration for struggling ninth graders in 34 U.S. public high schools, the Early College High Schools (ECHS) study conducted in 19 North Carolina public high schools, and the Small Schools of Choice (SSC) study conducted in 85 New York City public high schools.

[11]Online Appendix A presents supplementary tables for the present article, and Online Appendix B describes how we constructed our analysis samples and presents a detailed examination of sample attrition and treatment/control group pretest balance. Those results demonstrate that attrition rates for all studies are within or near the What Works Clearinghouse standard for "low attrition" (U.S. Department of Education, Institute of Education Sciences, 2017) and that treatment/control group pretest balance is excellent for all studies.

We selected outcome measures and pretests for each study in advance of its analysis. For outcome measures, we focused on standard academic indicators. When choosing among alternative outcome measures for a study, we tried to pick a limited number that: (1) were as comparable as possible across studies, (2) reflected a mix of mean treatment effects, (3) had acceptable attrition rates, and (4) for test scores, could be standardized as a broad-based z score reflecting the mean and standard deviation of scores for a meaningful population, like the nation (for HSIS, ASR, ASM, and ERO) a state (for ECHS), or a large urban school district (for SSCs).[12]

For pretest measures, we also chose academic test scores that could be standardized as broad-based z scores. Such z scores made it possible to compare the range of prior educational achievement for our analysis samples to that for meaningful populations. To limit the number of findings and their potential for multiple testing problems, we chose one pretest per study for all but ECHS and SSC. Because those two studies have data for two similar pretests, we used both pretests for a limited test of the sensitivity of our findings to the pretest used. For other studies with multiple pretests, we chose one pretest based on its attrition rate and our judgement about its likely ability to predict student outcomes and thus potentially predict treatment effects.

In terms of sample size, two of our studies (ASM and ASR) have analysis samples with roughly 2,000 students, three studies (HSIS, ERO and ECHS) have analysis samples with roughly 2,500–4,500 students, and one study (SSC) has an analysis sample with about 15,000 students for ninth-grade outcomes and 13,000 students for high school graduation rates.

Last, in terms of overall mean treatment effects, the studies and outcomes we chose reflect a mix of negligible effects (HSIS for one outcome, ASR for one outcome, and ERO for two outcomes), modest positive effects (ASM for one outcome and ERO for one outcome) and moderate to large positive effects (HSIS for two outcomes, ECHS for two outcomes, and SSCs for three outcomes).

Estimation

For each study/outcome/pretest scenario, we used two complementary approaches to explore the relationship between academic pretest scores and the effect of random assignment to an educational intervention on an educational outcome: non-parametric estimation and parametric estimation.

Our non-parametric approach stratifies sample members for a given scenario into three and five rank-ordered, equal-sized "bins" based on their pretest scores and then compares estimates of the mean treatment effect across bins. From sample data for each bin, we estimate its mean treatment effect as the coefficient, π, in the following regression model:

$$Y_i = \sum_j \alpha_j BLOCK_{ji} + \pi T_i + \gamma P_i + \sum_k \delta_k X_{ki} + e_i \quad (1)$$

[12]All test-score outcomes except the PPVT test of receptive vocabulary for the Head Start Impact Study could be standardized as a broad-based z score.

where

Y_i = the outcome for sample member i,
$BLOCK_{ji}$ = one if sample member i was from random-assignment block j and zero otherwise,[13]
T_i = one if sample member i was randomized to treatment and zero otherwise,
P_i = the pretest score for sample member i,
X_{ki} = the value of covariate k for sample member i,
e_i = an error term that varies independently and identically across sample members within experimental groups with a mean of zero and a variance that can differ for treatment and control-group members.

We then compared these estimates across bins and assessed the statistical significance of their variation based on the Hotelling T^2 statistic (Hotelling, 1951), hereafter referred to as Hotelling tests.

Although non-parametric estimation can provide important insights into how and how strongly treatment effects co-vary with pretest scores, it cannot by itself provide a useful *summary* of this relationship, which is essential for quantifying the likely generalizability of RDD treatment-effect estimates. Hence, for this purpose, our primary approach was parametric estimation of the following linear regression model for each scenario:

$$Y_i = \sum_j \alpha_j BLOCK_{ji} + \beta_1 T_i + \beta_2 T_i P_i + \gamma P_i + \sum_k \delta_k X_{ki} + e_i \tag{2}$$

where all variables are as defined as they were for Equation 1. This model specifies the following relationship between conditional mean treatment effects $\tau(P)$ and pretest scores (P):

$$\tau(P) = \beta_1 + \beta_2 P \tag{3}$$

Of primary interest is the slope, β_2, of Equation 3, which represents the average rate at which mean treatment-effects change per unit change in pretest scores across the range of pretest scores for a given scenario. Hence, it indicates the overall direction and strength of covariation between treatment effects and pretest scores. In addition, as described later, β_2 provides input to our predictions of RDD generalization bias.

As a secondary parametric approach for exploring the relationship between treatment effects and pretest scores, we estimated the following quadratic regression model:

$$Y_i = \sum_j \alpha_j BLOCK_{ji} + \beta_1' T_i + \beta_2' T_i P_i + \beta_3' T_i P_i^2 + \gamma P_i + \psi P_i^2 + \sum_k \delta_k X_{ki} + e_i \tag{4}$$

where all variables are defined as they were for Equations 1 and 2. This model specifies the following relationship between conditional mean treatment effects and pretests scores:

$$\tau(P) = \beta_1' + \beta_2' P + \beta_3' P^2. \tag{5}$$

[13]Randomized treatment assignment was blocked by a combination of site, student grade, and/or student cohort for the studies in our analysis.

Findings

The present section first describes our findings from each study to identify scenarios that do and do not exhibit treatment-effect/pretest-score covariation. For scenarios that do not exhibit such covariation, we argue that had a pretest-based RDD been used to estimate mean treatment effects, its results would have generalized well across the range of pretest scores examined and thus to relevant populations for the interventions being tested. For scenarios that exhibit covariation, we quantify the likely RDD generalization bias that it implies. As will be seen, our empirical findings reveal little evidence of substantial RDD generalization bias.

Individual Study Findings

Table 2 reports key linear and non-parametric findings for each study/outcome/pretest scenario we examined. The first column of findings in the table reports estimates of β_2 from Equation 2. The next column reports the statistical significance level (p value) for each estimate of β_2. The last two columns report p values for Hotelling tests of the statistical significance of observed variation in mean treatment effects across three bins and across five bins.

Now consider the findings, beginning with those for our youngest sample members, three- and four-year-olds in the national Head Start Impact Study. Results for this group focus on three outcomes. For two of those outcomes, letter-word identification and oral comprehension, we find no evidence of treatment-effect/pretest-score covariation. Their estimates of β_2 (a 0.017 and a –0.008 standard deviation change in the mean treatment effect per standard deviation increase in pretest scores, respectively) are quite small. Furthermore, Hotelling tests for these outcomes do not indicate statistically significant variation in mean treatment effects across three or five bins.

However, there is some evidence that Head Start assignment effects on receptive vocabulary co-vary with pretest scores. Specifically, the estimate of β_2 for this outcome implies a 0.059 standard deviation decline in mean treatment effects per standard deviation increase in pretest scores, and this estimate is statistically significant. That result is consistent with findings from the original Head Start Impact Study (Puma et al., 2010), which found compensatory program effects on receptive vocabulary, and subsequent research by Bloom and Weiland (2015), who argue that this phenomenon represents compensation for limited prior exposure to English because it only occurs for children from non-English speaking families. However, our Hotelling test results do not indicate statistically significant variation in treatment effects across three bins or across five bins.

Now consider the two after-school demonstration programs which served second- through fifth-graders at program centers from across the United States (25 centers per study). Findings in Table 2 provide no evidence of treatment-effect/pretest-score covariation for After-School Math. Its estimated value for β_2 is quite small (0.007). Also, Hotelling tests do not indicate statistically significant variation in mean treatment effects across three or five bins.

Findings for After-School Reading are mixed. Its estimated value of β_2 implies that mean treatment effects decline by 0.063 standard deviation per standard deviation increase in pretest scores, although this estimate is not quite statistically significant. On

Table 2. Summary of findings.

Study & outcome	Pretest	Linear Coefficient	Coefficient *p* value	Hotelling test *p* value	
				3 Bins	5 Bins
Head Start Impact Study					
Receptive vocab. score (*z*)	Letter/word score (*z*)	−0.059*	0.013	0.379	0.513
Letter/word score (*z*)	Letter/word score (*z*)	0.017	0.603	0.576	0.975
Oral comp. score (*z*)	Letter/word score (*z*)	−0.008	0.752	0.615	0.696
After School Math					
Math score (*z*)	Math score (*z*)	0.007	0.844	0.564	0.324
After School Reading					
Reading score (*z*)	Reading score (*z*)	−0.063	0.060	0.130	0.656
Enhanced Reading Opportunities					
% of required courses passed	Comp. score (*z*)	−0.738	0.253	0.116	0.196
Reading vocab. score (*z*)	Comp. score (*z*)	0.120*	0.017	0.017*	0.092
Reading comp. score (*z*)	Comp. score (*z*)	0.023	0.641	0.941	0.870
Early College High Schools					
9th grade on track (%)	Math score (*z*)	−5.509***	0.000	0.000***	0.000***
9th grade on track (%)	Reading score (*z*)	−4.176***	0.000	0.000***	0.000***
Graduated (%)	Math score (*z*)	−0.967	0.547	0.779	0.788
Graduated (%)	Reading score (*z*)	1.386	0.385	0.542	0.563
Small Schools of Choice					
9th grade credits accumulated (#)	Math score (*z*)	−0.104	0.270	0.354	0.292
9th grade credits accumulated (#)	Reading score (*z*)	−0.020	0.825	0.247	0.453
9th grade on track (%)	Math score (*z*)	0.899	0.420	0.617	0.999
9th grade on track (%)	Reading score (*z*)	0.125	0.907	0.299	0.453
Graduated (%)	Math score (*z*)	−0.103	0.930	0.226	0.306
Graduated (%)	Reading score (*z*)	−0.067	0.953	0.399	0.207

Source: The linear coefficient for each study/outcome/pretest scenario was estimated from data for each scenario's parametric analysis sample using Equation 2.
Note. A two-tailed test was used to assess the statistical significance of each estimated linear coefficient and a Hotelling test was used to test the statistical significance of differences across bins in their estimated mean treatment effects. Significance levels are indicated by *p* values and by ***≤ 0.001, **≤ 0.01, and *≤ 0.05.

the other hand, estimated mean treatment effects for each of five bins in Table 3 and each of three bins in Online Appendix Table A.1 do not exhibit a systematic pattern of variation across bins, and their Hotelling Test results do not indicate statistically significant treatment-effect variation.[14]

The next findings in Table 2 are for three outcomes of the Enhanced Reading Opportunity demonstration for ninth graders in 34 U.S. public high schools. Findings for two of these outcomes—reading comprehension scores and the percentage of required courses passed—do not indicate covariation between treatment effects and pretest scores.

However, reading-vocabulary scores exhibit some evidence that treatment effects covary with pretest scores. First, the estimated value of β_2 for this outcome implies a 0.120 standard deviation increase in mean treatment effects per standard deviation increase in pretest scores and is statistically significant. Second, Hotelling tests indicate statistically significant variation in mean treatment effects across three bins, although not across five bins. Third, findings in Online Appendix Table A.1 indicate that the

[14]Online Appendix Tables A.2 and A.3 report standard errors for our estimates of mean treatment effects for each of the three and five bins for each scenario.

Table 3. Non-parametric findings for five bins.

Study & outcome	Pretest	Mean treatment-effect estimate					Hotelling *p* value
		Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	
Head Start Impact Study							
Receptive vocab. score (*z*)	Letter/word score (*z*)	0.21***	0.12*	0.20**	0.09	0.11	0.513
Letter/word score (*z*)	Letter/word score (*z*)	0.24**	0.23**	0.21*	0.27**	0.19*	0.975
Oral comp. score (*z*)	Letter/word score (*z*)	0.00	0.03	0.01	−0.10	0.02	0.696
After School Math							
Math score (*z*)	Math score (*z*)	0.06	−0.03	0.14*	0.03	0.13*	0.324
After School Reading							
Reading score (*z*)	Reading score (*z*)	0.04	−0.03	0.03	−0.07	−0.03	0.656
Enhanced Reading Opportunities							
% of required courses passed	Comp. score (*z*)	1.12	1.31*	−0.24	0.06	0.95	0.196
Reading vocab. score (*z*)	Comp. score (*z*)	0.03	−0.07	0.00	0.08*	0.07	0.092
Reading comp. score (*z*)	Comp. score (*z*)	0.07	0.07	0.04	0.11**	0.06	0.870
Early College High Schools							
9th grade on track (%)	Math score (*z*)	16.15***	11.08***	3.44*	2.65*	0.50	0***
9th grade on track (%)	Reading score (*z*)	14.70***	8.06***	5.62***	3.21*	1.48	0***
Graduated (%)	Math score (*z*)	7.40	5.48	3.50	2.24	7.09*	0.788
Graduated (%)	Reading score (*z*)	−0.82	7.35	2.27	5.36	4.19	0.563
Small Schools of Choice							
9th grade credits accumulated (#)	Math score (*z*)	0.86***	0.94***	0.59***	0.98***	0.57***	0.292
9th grade credits accumulated (#)	Reading score (*z*)	0.70***	0.87***	1.02***	0.62***	0.61***	0.453
9th grade on track (%)	Math score (*z*)	8.39***	8.51***	7.90***	8.16***	7.68***	0.999
9th grade on track (%)	Reading score (*z*)	8.57***	8.01***	11.00***	4.97*	8.46***	0.453
Graduated (%)	Math score (*z*)	3.61	4.02	3.24	9.14***	6.27***	0.306
Graduated (%)	Reading score (*z*)	4.94	3.44	10.15***	3.23	6.10**	0.207

Source: The mean treatment effect for each bin within each study/outcome/pretest scenario was estimated using Equation 1 with data for that bin and scenario.

Note. A two-tailed test was used to assess the statistical significance of each bin's estimated mean treatment effect, and a Hotelling test was used to assess the statistical significance of differences across bins in their estimated mean treatment effects. Significance levels are indicated by *p* values and by ***≤ 0.001, **≤ 0.01, and *≤ 0.05.

mean treatment effect for bin three is appreciably larger than that for bin one or bin two in the three-bin findings.

Now, consider findings for the last two studies—Early College High Schools in North Carolina and Small Schools of Choice in New York City. Recall that these findings were replicated for two pretests to assess their sensitivity to the pretest used. Table 2 indicates that our findings are consistent across the two pretests for all outcomes for both studies.

However, results for the two Early College High School *outcomes* represent a stark contrast. On the one hand, Tables 2 and 3 provide no evidence of covariation between treatment effects on graduation rates and pretest scores. On the other hand, they provide strong evidence of strong covariation between treatment effects on ninth-grade on-track rates and pretest scores. For example, their estimates of β_2 indicate a 5.509 and 4.176 percentage-point decrease in the ECHS effect on ninth-grade on-track rates per standard deviation increase in math pretest scores and reading pretest scores, respectively, and these estimates are statistically significant. In addition, Hotelling tests indicate highly statistically significant treatment-effect variation across three and five bins. Furthermore, there are steep and consistent decreases in the magnitudes of estimated mean treatment effects across these bins (Table 3 and Online Appendix Table A.1, respectively).

To help understand this striking covariation, we examined the control-group pattern of counterfactual on-track rates for five and three bins (see Online Appendix Tables A.5

and A.4, respectively). Doing so revealed that ninth-grade on-track rates rapidly approach their maximum possible value of 100% as one moves across bins. For example, this rate increased from 74.0% to 87.0% to 96.1% to 97.3% to 98.6% across math pretest bins one through five and had a similar pattern for reading pretests. Thus, after bin two, there was almost no room for treatment to improve the outcome. Fortunately, a *ceiling effect* like this can be identified empirically (as we did here), which makes it possible to avoid making inappropriate RDD generalizations when it occurs.[15] Interestingly, no ceiling effect was observed for high school graduation rates, which are 70.1%, 81.8%, 87.4%, 88.3% and 86.0% for control-group members in math pretest bins one through five, respectively, and exhibit a similar pattern for reading pretests.

Now, consider the SSC findings in Table 2. As can be seen, these findings for three outcomes and two pretests provide no evidence of covariation between treatment effects and pretest scores. All estimates of β_2 are small in magnitude and are not statistically significant. Furthermore, no Hotelling test indicates statistically significant variation in treatment effects across bins.

As a final step, we identified scenarios with consistent evidence of a quadratic treatment-effect/pretest score relationship based on two criteria. The first criterion was whether the estimated quadratic coefficient, β'_3, in Equation 4 was statistically significant. If so, this suggests but does not fully demonstrate a non-monotonic treatment-effect/pretest-score relationship. The second criterion was whether *together* P^2 and P (the pretest score squared and the pretest score) in Equation 4 had a statistically significant ability to predict treatment effects. If the two criteria are met, this suggests a systematic overall relationship between treatment effects and pretest scores.[16]

Findings in Online Appendix Table C.4 indicate that only one of 18 scenarios met our two criteria—that for the relationship between graduation effects of Early College High Schools and their students' math pretest scores. (However, this scenario had no evidence of linear covariation between treatment effects and pretest scores.) We examine the bias implications of this finding and the preceding findings in the next section.

Implications for RDD Generalization Bias

On balance, findings in Tables 2, 3, and online Appendix Table A.1 provide evidence of linear covariation between treatment effects and pretest scores for four of the 18 scenarios we examined. In addition, findings in online Appendix C.4 provide evidence of quadratic covariation for one additional scenario. Hence, for 13 of the 18 scenarios we examined, we did not find evidence of covariation between treatment effects and pretests. This suggests that had an RDD been used to estimate treatment effects for those scenarios, the resulting findings would have generalized well across the range of pretest scores in those study samples and thus to relevant target populations for the

[15]A ceiling effect can be identified for positive impacts but not for negative effects.
[16]In our judgement, meeting the first criterion without meeting the second does not constitute consistent evidence of a non-monotonic relationship between treatment effects and pretest scores.

interventions being tested.[17] To help describe these populations and those for the other scenarios in our analysis, Table 4 reports the 10th, 50th, and 90th percentile pretest scores for each scenario plus the corresponding difference between their 90th and 10th percentile scores, which we refer to as an 80-percentile span.[18]

Note first that samples for five of the six studies in our analysis represent students with pretest scores that are disproportionately below their respective national, state, or district average. This is consistent with the compensatory nature of the interventions being tested. However, pretest scores for the ECHS sample are disproportionately above their state average, even though this intervention targets groups that historically have been underrepresented in college, such as students from specific racial/ethnic groups and from families with low incomes or no prior college attendance.

Note second that the 80-percentile pretest-score span is substantial for each study—with the smallest being 1.07 standard deviation for ERO, a highly targeted compensatory intervention, the largest being 2.73 standard deviations for HSIS, which has a nationally representative sample of program sites, and the remainder varying between 1.68 and 2.21 standard deviations. Because each of these samples represents a diverse target population that is policy relevant, a likely ability to generalize RDD findings to those target populations for 13 of 18 scenarios is quite promising.

Now consider the four scenarios which appear to exhibit some linear treatment-effect/pretest-score covariation. As noted earlier, the best summary of this covariation is the estimated value of β_2, which is the average rate at which treatment effects change per standard deviation change in pretest scores across the range of pretest scores examined. For example, the difference between mean Head Start effects on receptive vocabulary for children whose pretest scores differ by half a standard deviation—a potentially consequential pretest-score difference—is $\beta_2(0.5) = 0.059(0.5) \approx 0.030$ standard deviation, which is probably not a consequential treatment-effect difference for Head Start. The corresponding treatment-effect difference for children whose pretest scores differ by a full standard deviation is 0.059 standard deviation, which might or might not be considered consequential depending on the standard applied.

The preceding example illustrates a "*point-to-point*" treatment-effect comparison for two subpopulations with two different pretest scores. However, when assessing RDD generalizability, it is usually more relevant to compare the mean treatment effect at an RDD cut-point to the mean treatment effect for a target population defined by a range of pretest scores. For example, one might want to know how well an RDD estimate of the mean Head Start effect on receptive vocabulary would generalize to children with pretest scores between an RDD cut-point and one-half standard deviation below that cut-point. Thus, RDD generalizability is best framed as a "*point-to-range*" treatment-effect comparison.

[17]It is theoretically possible that with no overall impact/pretest-score covariation and no variation in mean impacts *across* pretest-score bins, that impact variation *within* bins could produce problematic RDD generalization bias. However, for this to occur would require: (1) a large and abrupt impact aberration within a bin, which seems unlikely, plus (2) an RDD cut-point that falls on this impact aberration, which seems unlikely. The joint occurrence of these two unlike conditions is thus very unlikely.

[18]These percentiles are for the non-parametric sample. For more information, Online Appendix Tables A.6 and A.7 report median pretest scores for each of three bins and each of five bins.

Table 4. Percentile pretest scores for each analysis sample.

Study & outcome	Pretest	10th Percentile score	50th Percentile score	90th Percentile score	80 Percentile span
Head Start Impact Study					
Receptive vocab. score (*z*)	Letter/word score (*z*)	−2.09	−0.74	0.65	2.73
Letter/word score (*z*)	Letter/word score (*z*)	−2.09	−0.74	0.65	2.73
Oral comp. score (*z*)	Letter/word score (*z*)	−2.09	−0.74	0.65	2.73
After School Math					
Math score (*z*)	Math score (*z*)	−1.64	−0.67	0.52	2.17
After School Reading					
Reading score (*z*)	Reading score (*z*)	−2.06	−1.08	−0.23	1.83
Enhanced Reading Opportunities					
% of required courses passed	Comp. score (*z*)	−1.60	−1.00	−0.53	1.07
Reading vocab. score (*z*)	Comp. score (*z*)	−1.60	−0.87	−0.53	1.07
Reading comp. score (*z*)	Comp. score (*z*)	−1.60	−0.87	−0.53	1.07
Early College High Schools					
9th grade on track (%)	Math score (*z*)	−0.83	0.35	1.38	2.21
9th grade on track (%)	Reading score (*z*)	−0.70	0.36	1.40	2.09
Graduated (%)	Math score (*z*)	−0.83	0.30	1.30	2.13
Graduated (%)	Reading score (*z*)	−0.77	0.35	1.38	2.15
Small Schools of Choice					
9th grade credits accumulated (#)	Math score (*z*)	−1.01	−0.13	0.68	1.69
9th grade credits accumulated (#)	Reading score (*z*)	−1.13	−0.30	0.71	1.84
9th grade on track (%)	Math score (*z*)	−1.05	−0.17	0.68	1.73
9th grade on track (%)	Reading score (*z*)	−1.17	−0.30	0.71	1.89
Graduated (%)	Math score (*z*)	−0.99	−0.13	0.68	1.68
Graduated (%)	Reading score (*z*)	−1.13	−0.23	0.71	1.84

Source: Findings in this table were computed from data for the non-parametric analysis sample for each study/outcome/pretest scenario. To facilitate interpretation, each pretest is expressed as a standardized *z* score based on the mean and standard deviation of scores on the same test for a broad-based reference population like the nation; a state; or a large, urban school district.

To provide such comparisons for the present analysis, we first defined alternative target-population pretest-score ranges. We next defined alternative target-population density functions across those pretest-score ranges. Then, for each combination of pretest-score range and density function, we predicted the difference between the mean treatment effect at one end of the pretest-score range (to emulate an RDD cut-point) and that for the full range (to emulate a target population). This predicted treatment-effect difference, $\Delta\tau$, is the bias, *GENBIAS*, we would expect if an estimate of the treatment effect at an RDD cut-point were generalized to the adjacent target population.

Our bias assessment is based on two illustrative target populations: (1) students with pretest scores between an RDD cut-point and one-half standard deviation below it, and (2) students with pretest scores between an RDD cut-point and a full standard deviation below it. Our assessment focuses on students below an RDD cut-point because most educational interventions are targeted on students who need assistance.[19] However, our approach and findings apply equally well to target populations that are above an RDD cut-point.

For each target population, we consider the three density functions in Figure 2. The top function in the figure is for densities that decrease at a constant rate from D_{max} at P_1 to zero at P_2. The middle function is a uniform distribution with density D_{max} for each pretest score from our RDD cut-point, P_1, to the far end of our target-population pretest-score range, P_2. The bottom function is for densities that increase at a constant rate from zero at P_1 to D_{max} at P_2. These simple density functions approximate a broad range of possible target-population distributions.

Note that a constantly decreasing density has most of its target population near the RDD cut-point, P_1. Thus, its mean treatment effect is heavily weighted toward treatment effects for population members who are close to the cut-point. We therefore refer to its *GENBIAS* findings as "optimistic" predictions.[20] At the opposite extreme, the constantly increasing density function has almost none of its target population near the RDD cut-point and its mean treatment effect is heavily weighted toward treatment effects for population members who are far from the cut-point. We refer to *GENBIAS* findings for this density function as "worst-case" predictions. Note, however, that for a pretest distribution with so few sample members near the cut-point, an RDD would be very underpowered and thus probably would not be conducted for other than extremely large samples. Between these two extremes, the uniform distribution gives equal weight to treatment effects that are close to and far from the cut-point. We thus refer to *GENBIAS* findings for this distribution as our "best-guess."

Online Appendix C derives the following *GENBIAS* expressions for a linear treatment effect model with each of our three density functions.

[19]Two important exceptions to this general treatment assignment tendency are the Thistlethwaite and Campbell (1960) study of the effects of National Merit Scholarships (which first introduced the regression discontinuity design) and the Angrist and Rokkanen (2015) study of Boston's elite exam schools discussed earlier.

[20]The three characterizations of our bias predictions discussed here hold for the full range of possible pretest scores when treatment effects are a monotonic function of pretest scores, as in our linear model. These characterizations also hold for potentially large portions of the pretest-score range when treatment effects are a non-monotonic function of pretest scores, as in our quadratic model, the results of which are discussed later.

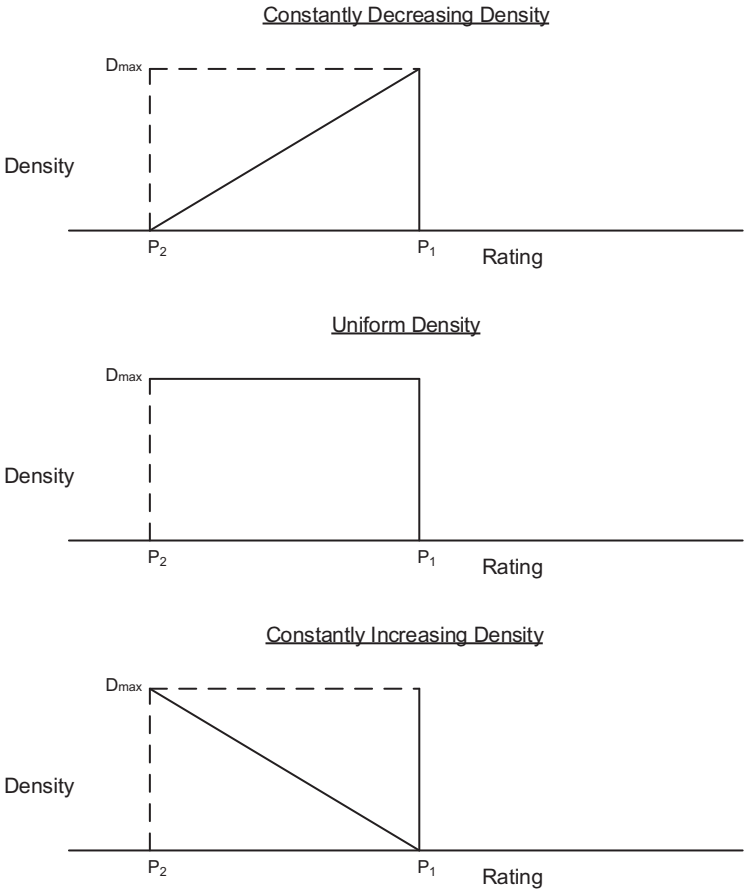

Figure 2. Alternative pretest-score density functions.

$$GENBIAS_{decreasing} = \frac{1}{3}\beta_2(P_1 - P_2) \text{ optimistic prediction} \quad (7)$$

$$GENBIAS_{uniform} = \frac{1}{2}\beta_2(P_1 - P_2) \text{ best guess} \quad (8)$$

$$GENBIAS_{increasing} = \frac{2}{3}\beta_2(P_1 - P_2) \text{ worst-case prediction} \quad (9)$$

Based on these expressions and our linear treatment-effect findings, Table 5 presents *GENBIAS* predictions for the four scenarios with evidence of linear treatment-effect/pretest-score covariation. Narrowing the focus to these four scenarios makes it clear that evidence of this covariation was only found for two types of outcomes: (1) vocabulary measures (for HSIS and ERO) and (2) a percentage measure with observations that are heavily concentrated near its maximum possible value (for ECHS). It is not clear why the two vocabulary measures exhibit treatment-effect/pretest-score covariation. However, as discussed earlier, covariation for the ECHS constrained percentage measure reflects a pronounced ceiling effect, which could be diagnosed in practice and therefore need not produce unwarranted RDD generalizations.

Now consider the magnitude of our linear bias predictions in Table 5. Note first that for a target population with a half–standard deviation pretest-score range, our best-guess

Table 5. Linear predictions of RDD generalization bias for scenarios with evidence of linear treatment-effect/pretest-score covariation.

Study & outcome	Pretest	Predicted bias		
		Optimistic prediction	Best guess	Worst-case prediction
		For a half–standard deviation rating range		
Head Start Impact Study				
Receptive vocab. score (*z*)	Letter/word score (*z*)	−0.010	−0.015	−0.020
Enhanced Reading Opportunities				
Reading vocab. score (*z*)	Comp. score (*z*)	0.020	0.030	0.040
Early College High Schools				
9th grade on track (%)	Math score (*z*)	−0.918	−1.377	−1.836
9th grade on track (%)	Reading score (*z*)	−0.696	−1.044	−1.392
		For a full–standard deviation rating range		
Head Start Impact Study				
Receptive vocab. score (*z*)	Letter/word score (*z*)	−0.020	−0.029	−0.039
Enhanced Reading Opportunities				
Reading vocab. score (*z*)	Comp. score (*z*)	0.040	0.060	0.080
Early College High Schools				
9th grade on track (%)	Math score (*z*)	−1.836	−2.755	−3.673
9th grade on track (%)	Reading score (*z*)	−1.392	−2.088	−2.784

Note. Optimistic predictions reflect a steeply declining density of individuals as ratings move away from their cut-point value; best-guess predictions reflect a uniform density of individuals as ratings move away from their cut-point value; worst-case predictions reflect a steeply rising density of individuals as ratings move away from their cut-point value.

biases are modest in magnitude ($\leq$0.030 standard deviation for the two vocabulary measures and $\leq$1.377 percentage points for the percentage measure). In addition, the worst-case biases (which, as noted earlier, are unlikely to occur in practice for an RDD) do not change this conclusion. For a target population with a full–standard deviation pretest-score range, our best-guess bias is modest in magnitude for receptive vocabulary (0.029 standard deviation) and moderate in magnitude for the other three measures (0.060 standard deviation for reading vocabulary and 2.088–2.755 percentage points for ninth-grade on-track rates). Here, too, the worst-case bias predictions do not change our overall conclusion. Furthermore, linear bias predictions reported in Online Appendix Tables C.2 and C.3 for the 14 scenarios without evidence of linear treatment-effect/pretest-score covariation indicate little margin for RDD generalization bias.

To assess the sensitivity of our linear *GENBIAS* predictions to the functional form of our model of the relationship between treatment effects and pretest scores, we reestimated those bias predictions for all 18 scenarios using the quadratic model described earlier. To do so, we first estimated $\beta_1', \beta_2',$ and β_3' from Equation 4 for each scenario (see Online Appendix Table C.4). Using these findings and quadratic *GENBIAS* expressions derived in Online Appendix C, we computed bias predictions for each scenario. Because quadratic *GENBIAS* predictions depend on the actual values of P_1 and P_2, not just on their difference, it was necessary to select a value of P_1 for each prediction. We set P_1 equal to the *mean sample pretest score* for each scenario because uncertainty about predictions from an estimated regression model is least at its sample mean. We then set P_2 for each scenario equal to one-half a standard deviation and a full standard deviation below P_1.[21] Online Appendix Tables C.5 and C.6 report the resulting *GENBIAS* findings.

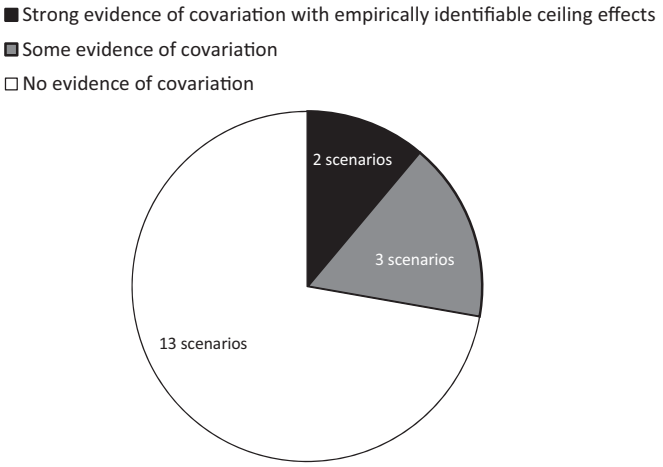

Figure 3. Pretest-score/treatment-effect covariation.

Now consider the single scenario that showed evidence of a quadratic treatment-effect/pretest score relationship—that for the relationship between graduation effects of Early College High Schools and their students' math pretest scores. Findings in Online Appendix Tables C.5 and C.6 for the scenario indicate RDD generalization bias that is modest in magnitude (ranging from 0.255–0.641 percentage points across our optimistic to worst-case predictions) for a half–standard deviation rating range and moderate in magnitude (ranging from 0.772–2.067 percentage points across our optimistic to worst-case predictions) for a full–standard deviation rating range. Thus, overall, our findings for a quadratic model indicate limited potential for substantial RDD generalization bias.

Discussion

To conclude, we briefly summarize and interpret the present findings, note their limitations and strengths, and consider several additional issues.

Summary and Interpretation

The present findings suggest that RDD estimates of educational treatment effects can often generalize to meaningful target populations. The first step toward this conclusion is the fact that little covariation was observed between educational intervention effects and students' pretest scores, which is the source of RDD generalization bias. For example, Figure 3 illustrates that, together, our parametric and non-parametric analyses provide no evidence of this covariation for 13 of the 18 study/outcome/pretest scenarios examined, some evidence of covariation for three scenarios, and strong evidence of covariation for only two scenarios. Furthermore, these latter two scenarios represent

[21]Sensitivity tests that we conducted indicate that our quadratic *GENBIAS* predictions for the one scenario with evidence of a quadratic relationship between treatment effects and pretest scores vary somewhat for moderate changes in the value of P_1.

pronounced ceiling effects that can be identified in practice, making it possible to prevent unwarranted RDD generalizations.

To help readers visualize the range of our best-guess RDD generalization bias findings, Figure 4 plots the 95% confidence interval for the best-guess linear bias estimate for each of our 18 scenarios in order from most negative to most positive bias. To report all bias estimates on the same scale, we converted those for the nine scenarios with outcomes measured in percentage points and those for the two scenarios with outcomes measured in course credits into effect sizes reported in standard deviation units. The top plot in Figure 4 is for a half–standard deviation target population, and the bottom plot is for a full–standard deviation target population. (Online Appendix Table A.8 reports the point estimate for each bias assessment).

To place these findings in context, note that when considering impact estimation bias due to sample attrition, The What Works Clearinghouse (WWC) states that, "A tolerable level of bias is defined as an effect size of 0.05 standard deviation or smaller . . . .The WWC's threshold for the tolerable level of bias was based on extensive consultation with experts" (U.S. Department of Education, Institute of Education Sciences, 2017, p. 10).[22] For easy reference, we superimposed this bias threshold on the plots in Figure 4.

Note that our best-guess linear bias estimates for a half–standard deviation target population vary from slightly negative (approximately –0.05 standard deviation) to slightly positive (0.03 standard deviation), with 15 of these 18 estimates and their confidence intervals clustering tightly around zero and none of them exceeding the WWC bias threshold. Furthermore, the two bias estimates with the largest magnitudes (approximately –0.05 and –0.04 standard deviation) represent the readily diagnosable ceiling effect discussed earlier for ECHS ninth-grade on-track rates. Corresponding findings for a full–standard deviation target population reflect the same pattern, although with magnitudes that are twice those of their counterparts for a half–standard deviation target population.[23] As can be seen, 15 of these estimates are within the WWC threshold for an acceptable bias, one is just outside the threshold, and the two that are furthest outside the threshold represent empirically identifiable ceiling effects. As a sensitivity test, for the one scenario with evidence of quadratic covariation between impacts and pretest scores, we computed the RDD generalization bias from our quadratic model. Our resulting best-guess quadratic bias estimates were –0.01 standard deviation for a half–standard deviation target population and –0.04 standard deviation for a full–standard deviation target population. Both estimates are within the WWC bias threshold.

Limitations and Strengths

When assessing the present results, it is important to consider the limitations and strengths of the analysis that produced them. Perhaps the most important limitation of

[22]WWC standards for acceptable bias from sample attrition should apply equally well to bias from RDD generalization, because both biases represent the distance between the expected value of an estimator and the actual value of its estimand, where an estimand represents a specific parameter for a specific population.

[23]It is interesting to note that two of the three previous studies of RDD generalizability that we discussed earlier (Angrist & Rokkanen, 2015; Wing & Cook, 2013) found empirical support for RDD generalizability, whereas the third study (Dong & Lewbel, 2015) found evidence of RDD generalization bias.

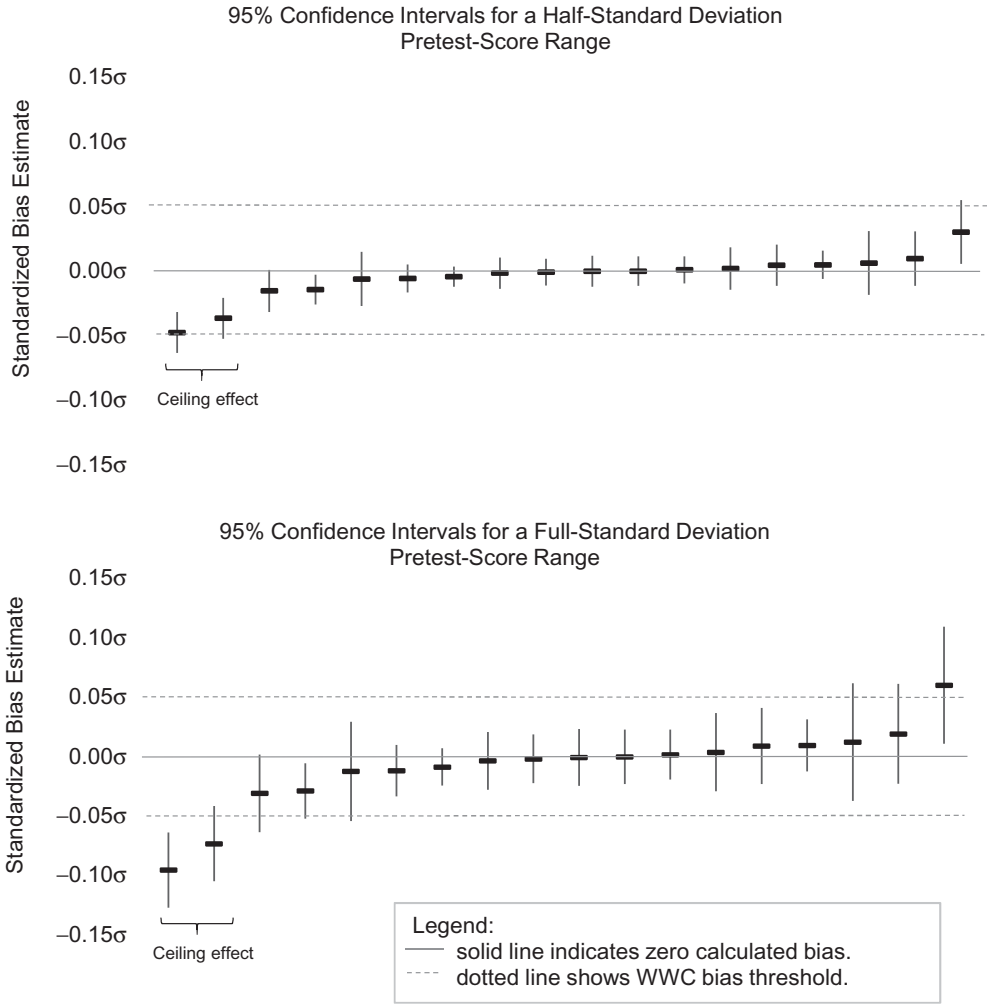

Figure 4. Linear best-guess bias in standard deviation units. Source: Bias estimates in this figure were obtained as explained in Online Appendix C and then converted into standard deviation units where necessary. Point estimates are shown in Online Table A.8. Note: the What Works Clearinghouse (WWC) states that "A tolerable level of bias is defined as an effect size of 0.05 standard deviations or smaller . . . .The WWC's threshold for the tolerable level of bias was based on extensive consultation with experts" (U.S. Department of Education, Institute of Education Sciences, 2017, p. 10).

this analysis is that it is based on the experience of only six RCTs. Thus, *its* generalizability is limited. Fortunately, those six RCTs span the developmental gamut from preschool to high school and involve large and diverse samples from many locations across the United States. Hence, they reflect a wide range of educational practices, educational environments, and student populations. In addition, they reflect the experience of six quite different interventions.

A second potential limitation of the present analysis is its inability to identify with confidence precise functional forms for relationships between treatment effects and pretest scores. To minimize this threat to the validity of our findings, we took several steps. First, we conducted a non-parametric analysis that provides a simple, direct and

informative *look* at treatment-effect/pretest-score covariation *patterns* unconstrained by assumed functional forms.

Second, we conducted two alternative parametric analyses that together represent the most frequently cited theories of relationships between educational treatment effects and student pretest scores. Our linear parametric analysis imposes a monotonic relationship that can represent either a compensatory theory of treatment effects (e.g., Ramey & Ramey, 1998) with negative covariation between treatment effects and pretest scores, or a "skills begets skills" theory of treatment effects (e.g., Cunha & Heckman, 2007) with positive covariation. In addition, our quadratic parametric analysis imposes a non-monotonic relationship that can represent a "bubble-kid" or "sweet-spot" theory with the largest positive effects occurring for students with midrange pretest scores, who might have a strong enough educational base to effectively participate in an intervention but enough room for growth to exhibit demonstrable improvement. We are not aware of established theories that predict more complex functional forms with multiple local maximum or minimum treatment effects. However, if such cases arise in future research of the type we have presented, researchers might want to consider using more flexible (although complex) data-driven estimation methods that are emerging in the literature, such as Bayesian Additive Regression Trees (Hill, 2011).

Additional Issues

Before concluding our discussion, it is useful to consider four additional issues. First is the role of *pretest reliability* in producing the present dearth of observed treatment-effect/pretest-score covariation. As described earlier, covariation between treatment effects and an empirical measure (e.g., student pretest scores) is the product of two factors: (1) the reliability of the measure and (2) covariation between treatment effects and whatever is being measured systematically. Thus, in theory, one explanation for the present dearth of observed covariation is weak pretest reliability. To explore this possibility, we reviewed the literature on the reliability of the pretests used for the present analysis (see Online Appendix D). Findings from our review suggest that the reliability of those pretests is uniformly high, ranging mainly from about 0.85–0.95 (which is probably because we used well-established pretests). Hence, weak pretest reliability cannot explain the weak treatment-effect/pretest covariation that we found.[24] Instead, these findings probably represent weak covariation between treatment effects and true prior academic achievement.

A second issue to consider is the use of *sample* density functions to help assess RDD generalizability in practice.[25] Specifically, for any given RDD study we recommend that researchers examine the location of the cut-point in the sample density distribution. If the cut-point is in a high-density region, the proportion of sample members with ratings near the cut-point—and thus with potentially similar treatment effects—is larger than if the cut-point is in a low-density region. Thus, to the extent that a study sample represents a relevant target population, RDD findings are more likely to generalize to a

[24]Because the reliability estimates we found were based on different reliability measures for different pretests (e.g., internal consistency, split-half or test-retest reliability), these findings are not fully comparable.
[25]We thank Mike Weiss for this suggestion.

relevant population when their cut-points are in a high-density sample region than when their cut-points are in a low-density sample region, other things being equal.

A third issue to note is that all RDD generalization bias estimates that we report are for target populations that lie on *one side* of a potential RDD cut-point. However, it is likely that policy-relevant target populations for some educational interventions *straddle* an RDD cut-point. For example, it might be important to estimate the mean effect of an educational intervention for a target population that lies in the pretest-score neighborhood of an RDD cut-point (both above and below it). In such cases, if the relationship between intervention effects and pretest scores is approximately monotonic, the mean intervention effect at the cut-point might be very close to the mean intervention effect for the target population, even with strong covariation between intervention effects and pretest scores. This is because a negative bias on one side of the cut-point will tend to offset the corresponding positive bias on the other side of the cut-point. For such target populations, the present findings probably overstate RDD generalization bias, perhaps by a lot. Indeed, for a symmetric target-population density function that is centered on an RDD cut-point and a linear relationship between intervention effects and pretest scores, RDD generalization bias for the target population will be zero, regardless of the strength or direction of covariation between intervention effects and pretests.

A fourth and final issue to consider is the relative roles of the present approach to RDD generalizability and those of the three approaches discussed earlier. For example, recall that Wing and Cook (2013) propose a pretest-enhanced RDD for extrapolating RDD findings based on observable relationships between a pretest and post-test measured on the same scale by the same instrument and a rating which measures something different. Although promising, this approach is limited to situations with the requisite pretests, outcomes, and ratings.

Next, recall that Angrist and Rokkanen (2015) propose a method for extrapolating RDD findings by conditioning on covariates that render treatment assignment conditionally ignorable. In addition, they demonstrate how to test the validity of this conditional ignorability assumption (their CIA) for a given RDD. If passed, this test opens the door to using conventional regression or matching methods to estimate treatment effects away from an RDD cut-point. Although promising, this approach requires covariates that predict RDD treatment assignment and individuals' potential outcomes well enough to eliminate the conditional correlation between potential outcomes and treatment assignment.

Last, recall that Dong and Lewbel (2015) propose an approach for using standard RDD methods to estimate the treatment-effect change rate at an RDD cut-point to explore the causal implications of very small changes in a treatment-assignment threshold. Although promising, this approach is not readily applicable for estimating the mean treatment effect for a target population with ratings that are not immediately adjacent to an RDD cut-point.

In contrast to the preceding approaches which use data from a given RDD study to help generalize treatment-effect estimates from that study, the present approach uses RCT data from a group of studies to assess the likely generalizability of RDD estimates from other studies. Thus, although the present approach can help to inform

expectations about the generalizability of specific RDD findings, it cannot directly assess or enhance the generalizability of those findings.

Each of the preceding approaches attempts to address different questions about RDD generalizability, each approach is applicable to different situations, and each approach has different strengths and weaknesses. However, together as a methodological portfolio, the four approaches comprise a formidable set of tools for expanding and enhancing the use and usefulness of the regression discontinuity design, which in the judgement of many researchers (e.g., Lee & Lemieux, 2010), is an especially strong quasi-experimental design.

Acknowledgments

The authors thank Michael Weiss, Pei Zhu, Kristin Porter, Himani Gupta, Daniel Cullinan and Alec Gilfillan for their valuable input.

Funding

This article was supported by grant RD305D140012 from the Institute of Education Sciences and grant 201500035 from the Spencer Foundation. However, all views and information presented are the sole responsibility of the authors.

ARTICLE HISTORY

Received 10 October 2018
Revised 12 May 2019
Accepted 14 May 2019

References

Angrist, J. D., & Pischke, J. S. (2015). *Mastering metrics: The path from cause to effect*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331–1344.

Black, A. R., Somers, M. A., Doolittle, F., Unterman, R., & Grossman, J. B. (2009). *The evaluation of enhanced academic instruction in after-school programs: Final report* (NCEE 2009–4077). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43–82. doi:10.1080/19345747.2011.578707

Bloom, H. S., & Michalopoulos, C. (2011). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects on subgroups. *Prevention Science*, *14*(2), 179–188.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842. doi:10.1080/19345747.2016.1264518

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877. doi:10.1080/19345747/2016.1271069

Bloom, H. S., & Unterman, R. (2014). Can small high schools of choice improve educational prospects for disadvantaged students? *Journal of Policy Analysis and Management*, *33*(2), 290–319. doi:10.1002/pam.21748

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study (MDRC Working Papers on Research Methodology). Retrieved from https://www.mdrc.org/sites/default/files/quantifying_variation_in_head_start.pdf

Booher-Jennings, J. (2005). Below the bubble: Educational triage and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231–268. doi:10.3102/00028312042002231

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brown, R. S., & Dale, S. B. (2007). The research design and methodological issues for the cash and counselling evaluation. *Health Services Research*, *42*(1Pt2), 414–445.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, *37*(2), 403–429.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Cooper, H. M., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47. doi:10.1257/aer.97.2.31

Dong, Y., & Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *The Review of Economics and Statistics*, *97*(5), 1081–1092.

Edmunds, J., Unlu, F., Glennie, E., Bernstein, L., Fesler, L., Furey, J., & Arshavsky, N. (2017). Smoothing the transition to postsecondary education: The impact of the early college model. *Journal of Research on Educational Effectiveness*, *10*(2), 297–325. doi:10.1080/19345747.2016.1191574

Goodman, J. (2008). Who merits financial aid? Massachusetts' Adams Scholarship. *Journal of Public Economics*, *92*(10–11), 2121–2131. doi:10.1016/j.jpubeco.2008.03.009

Hahn, J., Todd, P., & Van Der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica*, *69*(1), 201–209. doi:10.1111/1468-0262.00183

Hill, J. (2011). Bayesian non-parametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217. doi:10.1198/jcgs.2010.08162

Hotelling, H. (1951). *A generalized T test and measure of multivariate dispersion*. Chapel Hill, NC: University of North Carolina.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. doi:10.1016/j.jeconom.2007.05.001

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. Senate House elections. *Journal of Econometrics*, *142*(2), 807–828. doi:10.1016/j.jeconom.2007.05.004

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355. doi:10.1257/jel.48.2.281

Nomi, T., & Raudenbush, S. W. (2016). Making a success of "Algebra for All": The impact of extended instructional time and classroom peer skill in Chicago. *Educational Evaluation and Policy Analysis*, *38*(2), 431–451. doi:10.3102/0162373716643756

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill Book Company.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., & Friedman, J. (2010). *Head Start Impact Study final report*. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_final.pdf

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist*, *53*(2), 109–120. doi:10.1037/0003-066X.53.2.109
Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from cross-site distributions of program impacts using multi-site trials. *American Journal of Evaluation*, *36*(4), 475. doi:10.1177/1098214015600515
Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *The Lancet*, *365*(9454), 176–186. doi:10.1016/S0140-6736(05)17709-5
Somers, M.-A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The Enhanced Reading Opportunities Study Final Report: The impact of supplemental literacy courses for struggling ninth-grade readers* (NCEE 2010–4021). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. doi:10.1037/h0044319
Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. In P. J. Lavrakas (Ed.), *Experimental methods in survey research: Techniques that combine random sampling with random assignment*. New York, NY: Wiley.
U.S. Department of Education. (2017). Institute of Education Sciences, What Works Clearinghouse. *Standards Handbook, version 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.
Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. doi:10.1080/19345747.2017.13000719
Wing, C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparision. *Journal of Policy Analysis and Management*, *32*(4), 853–877. doi:10.1002/pam.21721
Wooldridge, J. F. (2009). *Introductory econometrics: A modern approach* (4th ed.). Mason, OH: South-Western CENGAGE Learning.

Appendix

Study synopses

Head Start Impact Study (HSIS)

Intervention: The Head Start (HS) program seeks to improve school readiness among children from low-income families. The most common HS programs are center-based programs, engaging children primarily in a classroom setting and providing at least two home visits per year. Other HS models include home-based programs, family child care programs, and combination programs. HS participation can last for up to two years.

Target population: Low-income children (3- to 4-year-olds) in a nationally representative sample of HS programs, excluding those intended to serve specific populations.

Study design: Individual random assignment within HS centers. The study compares students who were offered enrollment in HS with students who were not allowed to enroll in HS.

Outcomes: Cognitive measures from an abbreviated version of the Peabody Picture Vocabulary Test-III, and the letter-word identification, oral comprehension, and applied problems subscales of the Woodcock-Johnson III. Socioemotional measures created based on parent-reported items from the Child Behavior Checklist and the Leiter-R Assessor Report. All assessments were done at the end of the year in which the student enrolled in Head Start.

Study sample: Approximately 300 HS centers and 3,500 children.

Report: Bloom and Weiland (2015); Puma et al. (2010).

After-School Math (ASM)

Intervention: A structured and academically rigorous after-school program using the Harcourt math curriculum. Program enrollment could last for up to two years.

Target population: Students in Grades 2–5 who attend after-school programs.

Study design: Individual random assignment within after-school center/grade/cohort blocks. The study compares students who were randomized to an academically oriented after-school program with students who were randomized to less formal academic support offered in a regular after-school program. Both programs operated in the same after-school centers.

Outcome: SAT-10 total math score at the end of students' first year in the study.

Study sample: 25 after-school centers and approximately 2,500 students (Black, Somers, Doolittle, Unterman, & Grossman, 2009).

Report: Black et al. (2009).

After-School Reading (ASR)

Intervention: A structured and academically rigorous after-school program using the Success for All reading curriculum. Program enrollment could last for up to two years.

Target population: Students in Grades 2–5 who attend after-school programs.

Study design: Individual random assignment within after-school center/grade/cohort blocks. The study compares students who were randomized to an academically oriented after-school program with students who were randomized to less formal academic support offered in a regular after-school program. Both programs operated in the same after-school centers.

Outcome: SAT-10 total reading score at the end of students' first year in the study.

Study sample: 25 after-school centers and approximately 2,300 students (Black et al., 2009). These centers and students differ from those in the After-School Math demonstration.

Report: Black et al. (2009).

Enhanced Reading Opportunities (ERO)

Intervention: Ninth-grade students take a supplemental reading course in place of an elective class, using either the Reading Apprenticeship Academic Literacy (RAAL) program or the Xtreme Reading program. Program enrollment was intended to last for one year.

Target population: Ninth-grade students whose reading ability is at least two years below grade level.

Study design: Individual random assignment within each school by cohort block. In addition, schools were randomly assigned to one of the two reading curricula. The study compares students who were selected to enroll in the supplemental reading class (using either RAAL or Xtreme Reading) with students who took some other elective class regularly offered by the school.

Outcomes: Reading comprehension and vocabulary scores on the GRADE assessment, and credits earned as a percentage of credits required for graduation during the program year and follow-up year.

Study sample: 34 public schools and approximately 5,500 students (Somers et al., 2010).

Report: Somers et al. (2010).

Early College High School (ECHS)

Intervention: Early College High Schools provide students with concurrent high school and college experiences. Students attend high school on a college campus, enroll in college courses,

and are expected to complete two years of transferable college credits or an associate degree by the time they earn their high school diploma.

Target population: High school students who are underrepresented in college: first in their family to go to college; low-income students; members of racial and ethnic groups that are underrepresented.

Study design: Schools were selected based on whether they were overenrolled and agreed to use a lottery system to assign students. Students were randomly assigned within each school by cohort block. In some cases, the lottery for a given school by cohort block gave students different probabilities of selection into the program group.

Outcomes: Ninth-grade "on-track" indicator and five-year graduation rate.

Study sample: 19 schools and approximately 4,000 students (depending on the outcome).

Report: Edmunds et al. (2017).

Small Schools of Choice (SSCs)

Program: Small schools of choice are academically nonselective high schools with typically around 500 students, and they are based on the principles of academic rigor, real-world relevance, and personalized relationships.

Target population: Incoming first-time ninth graders in New York City who listed an SSC among their high school choices.

Study design: Individual random assignment from a lottery-based admission system. The study compares students who indicated interest in enrolling in an SSC and were assigned by lottery to an SSC with those who indicated interest in an SSC but were not assigned to one.

Outcomes: The number of course credits accumulated in ninth grade, the percentage of first-time ninth graders who were on track toward graduation, and the percentage of first-time ninth graders who graduated from high school in four years.

Study sample: 85 SSCs and approximately 13,000 to 15,000 students (depending on the outcome).

Report: Bloom and Unterman (2014).

Notes. Adapted with permission from Weiss et al. (2017).