Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

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

Asymdystopia: The Threat of Small Biases in Evaluations of Education Interventions That Need to Be Powered to Detect Small Impacts

John Deke, Thomas Wei & Tim Kautz

To cite this article: John Deke, Thomas Wei & Tim Kautz (2021) Asymdystopia: The Threat of Small Biases in Evaluations of Education Interventions That Need to Be Powered to Detect Small Impacts, Journal of Research on Educational Effectiveness, 14:1, 207-240, DOI: 10.1080/19345747.2020.1849480

To link to this article: https://doi.org/10.1080/19345747.2020.1849480

Published online: 16 Apr 2021.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2021, VOL. 14, NO. 1, 207–240
https://doi.org/10.1080/19345747.2020.1849480

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

Asymdystopia: The Threat of Small Biases in Evaluations of Education Interventions That Need to Be Powered to Detect Small Impacts

John Deke[a], Thomas Wei[b] and Tim Kautz[a]

[a]Mathematica, Inc, Princeton, New Jersey, USA; [b]Institute of Education Sciences, New York, New York, USA

ABSTRACT

Evaluators of education interventions are increasingly designing studies to detect impacts much smaller than the 0.20 standard deviations that Cohen characterized as "small." While the need to detect smaller impacts is based on compelling arguments that such impacts are substantively meaningful, the drive to detect smaller impacts may create a new challenge for researchers: the need to guard against smaller biases. The purpose of this article is twofold. First, we examine the potential for small biases to increase the risk of making false inferences as studies are powered to detect smaller impacts, a phenomenon we refer to as *asymdystopia*. We examine this potential for two of the most rigorous designs commonly used in education research—randomized controlled trials and regression discontinuity designs. Second, we recommend strategies researchers can use to avoid or mitigate these biases.

ARTICLE HISTORY
Received 4 December 2019
Revised 10 September 2020
Accepted 13 September 2020

KEYWORDS
Bias; attrition; regression discontinuity designs; randomized controlled trials

Introduction

Evaluators of education interventions increasingly need to design studies to detect impacts much smaller than the 0.20 standard deviations that Cohen (1988) characterized as "small." For example, an evaluation of Response to Intervention from the Institute of Education Sciences (IES) detected impacts ranging from 0.13 to 0.17 standard deviations (Balu et al., 2015), and IES' evaluation of the Teacher Incentive Fund detected impacts of just 0.03 standard deviations (Chiang et al., 2015).

The drive to detect smaller impacts is in response to strong arguments that, in many contexts, impacts once deemed "small" can still be meaningful (Kane, 2015). Hill et al. (2008) and Lipsey et al. (2012) suggest multiple substantive benchmarks for assessing what a "meaningful" impact would be for a given intervention and context. These benchmarks often suggest that impacts less than 0.20 standard deviations are meaningful. For example, under the cost-effectiveness benchmark, smaller impacts may be deemed meaningful when evaluating less-expensive interventions.

CONTACT John Deke jdeke@mathematica-mpr.com Mathematica, Inc, Princeton, NJ, USA.

Though based on a compelling rationale, the drive to detect smaller impacts may create a new challenge for researchers: the need to guard against relatively smaller biases. When studies were designed to detect impacts of 0.20 standard deviations or larger, it may have been reasonable for researchers to regard small biases as ignorable. For example, a bias of 0.03 standard deviations might have been ignorable in a study that could only detect an impact of 0.20 standard deviations. But for a study designed to detect much smaller impacts, such as Chiang et al. (2015) in which the impact estimate was 0.03 standard deviations, a bias of 0.03 standard deviations is no longer small—it is enormous.

The purpose of this article is twofold. First, we examine the potential for small biases to increase the risk of making false inferences (specifically, the rate at which a null hypothesis of zero impact is rejected under the null hypothesis significance testing [NHST] framework[1]) as studies are powered to detect smaller impacts. We refer to this phenomenon as *asymdystopia*.[2] Second, we recommend strategies researchers can use to avoid or mitigate these biases.

This article examines the potential for asymdystopia in two common education research designs—randomized controlled trials (RCTs) and regression discontinuity designs (RDDs). When perfectly executed, both designs produce valid causal impact estimates, but studies that use these designs typically have flaws that can be a source of bias. One of the more ubiquitous flaws, in theory and in practice, is attrition in RCTs and regression misspecification in RDDs. We thus focus on these two common sources of bias and examine whether they become increasingly problematic when studies are designed to detect smaller impacts. However, asymdystopia is likely not limited to these two specific biases but rather is a more general phenomenon, as we briefly illustrate with a few additional examples later in the article.

More specifically, we address two primary research questions:

1. **How problematic is attrition bias in RCTs when studies are powered to detect smaller impacts?** We explore this question using an attrition model for RCTs that is used in several federal evidence reviews. This model assumes that attrition bias is ignorable as long as it accounts for less than 20% of whatever size impact is deemed substantively important. Using this model and data on attrition from past studies, we examine three key issues. First, we consider how attrition may become less acceptable, leading to higher rates of false inferences, as studies are powered to detect smaller effects. Second, we discuss contexts in which more favorable assumptions about the relationship among attrition, outcomes, and treatment status may allow for greater tolerance of attrition, even in studies that are powered to detect small effects. Third, we provide evidence on the feasibility

[1]We acknowledge the recent critiques of the NHST framework (Amrhein et al., 2019; Wasserstein & Lazar, 2016) and do not intend for this article to implicitly endorse its continued use. The issues raised in this article are equally applicable to any inferential method (for example, Bayesian posterior probabilities) that ignores small biases when assessing and reporting uncertainty in studies powered to detect small impacts. We therefore stick with the NHST framework in this article for simplicity and because it is likely to be familiar to the widest range of readers.

[2]Some studies—particularly retrospective nonexperimental studies using administrative data—have the statistical power to detect effects that are too small to be substantively important. This article does not focus on "overpowered" studies. Instead, we focus on studies that are designed to have just enough statistical power to detect the smallest impact that is substantively important.

of achieving lower attrition rates in future studies that are powered to detect small impacts, based on an analysis of attrition in past RCTs.

2. **How problematic is functional form misspecification bias in RDDs when studies are powered to detect smaller impacts?** In an RDD study, treatment and comparison groups are formed using a cutoff on a continuous assignment variable.[3] Researchers must account for differences in the assignment variable between the treatment and comparison groups when estimating RDD impacts. For example, suppose a cutoff on a math test is used to assign students to an intervention that provides after-school help on homework. Students below the cutoff are in the treatment group, and students above the cutoff are in the comparison group. When estimating impacts, researchers regression-adjust for the fact that students in the treatment group are lower math achievers to begin with. If the functional form for this regression is incorrect (e.g., specifying a linear relationship when the true relationship is nonlinear), then the estimated impact could be biased. Increasing a study's sample size decreases this bias while also increasing the precision of the impact estimate—typically a win-win situation. However, if the precision increases faster than the bias decreases, it becomes relatively more likely for a biased impact estimate to be statistically significant, thereby increasing the risk of making a false inference. We use Monte Carlo simulations to assess what happens as the sample size of the RDD increases under varying assumptions regarding the true functional form. Specifically, we examine the effect of a larger sample size on statistical power, functional form misspecification bias, and the accuracy of estimated p-values (or confidence intervals) when using current state-of-practice data-driven bandwidth selectors proposed by Imbens and Kalyanaraman (2012) and by Calonico et al. (2014). We also verify that a popular method proposed by Calonico et al. (2014) to better account for misspecification bias, indeed reduces false inferences in our education contexts.

Across both investigations, our findings suggest that biases that might have once been reasonably ignorable can pose a real threat in evaluations that are powered to detect small impacts. Our article identifies and quantifies some of these biases and shows that they are important to consider when designing evaluations and when analyzing and interpreting evaluation findings. Our findings should *not* be interpreted as suggesting that researchers should avoid powering evaluations to detect small impacts. The problem of small biases is real but surmountable—so long as it is not ignored.

The remainder of the article is organized into four sections. The next section motivates the need to detect smaller impacts and how this can lead to asymdystopia. The following two sections present the methods and findings for our two primary research questions. The final section concludes with a discussion of implications and recommendations for researchers.

[3]The term "assignment variable" is often used interchangeably with "forcing variable," "running variable," and "score." Truly continuous assignment variables are atypical in practice, although methods do exist to account for discreteness (for example, Armstrong & Kolesár, 2018; Barreca et al., 2016; Kolesár & Rothe, 2018; Lee & Card, 2008).

The Power to Detect Small Impacts and the Potential for Asymdystopia

Ideally, evaluations would be designed so that their minimum detectable effect (MDE) is calibrated to be the same as the smallest *substantively important impact*. An impact is "detected" if it is *statistically significant*—that is, if the estimated impact is of a magnitude that is very unlikely to occur when the true impact is zero. To detect smaller impacts with high probability, an evaluation typically needs a larger sample size. Because larger sample sizes lead to higher evaluation costs, researchers and funders typically seek to design studies that are just large enough to detect a substantively important impact.[4]

In his seminal book, Cohen (1988) suggested three thresholds researchers can use as a general guide for whether an impact is substantively important or "meaningful." He suggested that impacts ranging from 0.20 to 0.49 standard deviations are meaningful but "small." Impacts larger than 0.50 are "medium," and those exceeding 0.80 are "large." Cohen acknowledged that he based these thresholds on his own subjective judgments and advised caution in how they are applied. Still, the thresholds have been widely cited (Lipsey et al., 2012) and have served as benchmarks for some time in a number of fields, including education. For example, until recently, the What Works Clearinghouse (WWC) had long defined a "substantively important" impact to be *at least* 0.25 standard deviations (What Works Clearinghouse [WWC], 2008, 2020).[5] Similarly, many older IES evaluations were designed to detect impacts in the range of 0.20–0.25 (Agodini & Harris, 2010; James-Burdumy et al., 2012, 2008).

Some researchers have argued more recently that using Cohen's benchmarks to design evaluations in education is often difficult to justify. Hill et al. (2008) and Lipsey et al. (2012) suggest a range of benchmarks for assessing what a "meaningful" impact would be for a given intervention in a given context. The benchmarks include how an impact of an intervention compares to typical annual growth in student outcomes; policy-relevant performance gaps between types of students (e.g., between black students and white students); observed impacts in similar contexts; and the impact relative to cost so that less expensive interventions would require smaller impacts to be meaningful.

In addition, smaller impacts might be substantively meaningful for secondary outcomes that the intervention affects less directly. Many interventions are designed to have a large impact on a *proximal* outcome that is closely aligned to the intervention and often measured shortly after the end of the intervention. For example, a study of an after-school program offering help on homework might examine impacts on homework completion rates. Policymakers, however, might also be interested in *distal* outcomes that the intervention targeted less directly but could still be impacted. Continuing the example, improvements in homework completion might ultimately lead to gains on state achievement tests. The impact on distal outcomes is likely to be smaller than the impact

[4]See Bloom (2005); Bloom et al. (2007); Deke and Dragoset (2012); Hedges and Hedberg (2007); Murray (1998); and Schochet (2008a, 2008b) for more information about calculating statistical power in both RCTs and RDDs.

[5]The What Works Clearinghouse, managed by the U.S. Department of Education's Institute of Education Sciences, systematically reviews and synthesizes education research studies with the goal of providing a reliable source of scientific evidence for what works in education to improve student outcomes. For more information, see http://ies.ed.gov/ncee/wwc/.

on proximal outcomes because distal outcomes are influenced by a wider range of factors that are beyond the scope of the intervention to influence.

Perhaps reflecting these considerations, more recent education evaluations have sought to detect impacts on distal outcomes much smaller than 0.20 standard deviations, such as Balu et al. (2015) and Chiang et al. (2015) that detected impacts as small as 0.13 and 0.03 standard deviations. Requests for proposals to conduct IES evaluations in recent years have similarly asked offerors to detect impacts on student achievement as small as 0.10 (e.g., the Impact Study of Feedback for Teachers Based on Classroom Videos and the Impact Evaluation to Inform the Teacher and School Leader Incentive Program).

The Potential for Asymdystopia

Asymptopia has been described as a place where "data are unlimited and estimates are consistent" (Leamer, 2010). An estimate is consistent if the expected value of the estimate approaches the true value of the parameter being estimated as the sample size approaches infinity. Impact estimates from RCTs and RDDs are both consistent if all the assumptions underpinning the methods are satisfied. Of course, asymptopia can never be achieved because data are never unlimited. Every study has a finite sample size. Nevertheless, it is tempting to believe that having more data will always lead us closer to the correct answer and reduce the probability that we draw false inferences.

We define *asymdystopia* as a context in which a larger (but finite) sample size is not necessarily better and could even be worse from the perspective of controlling the Type 1 error rate. There has historically been a strong aversion to falsely concluding that an intervention works when in fact it does not. Researchers therefore typically prefer to limit the occurrence of Type 1 errors to 5% by only declaring an impact statistically significant if the *p*-value is 0.05 or less (or, equivalently, if the magnitude of the t-statistic exceeds an appropriate "critical value"). But if, as a study becomes larger, the standard error of the impact estimate shrinks while bias stays the same (or shrinks less than the standard error), then Type 1 errors could become more common. This is because the denominator of the t-statistic (the standard error) is shrinking faster than the numerator (the biased point estimate).

Figure 1 provides an illustrative example of asymdystopia. In this example, the true impact is zero, as represented by the dashed line. Each dot represents the estimated impact for a given sample size, which is equivalent to the bias in this example because the true impact is zero. The bars above and below the dots represent 95% confidence intervals, so that an impact is statistically significant if the interval does not include zero. At a relatively small sample size ($N = 200$), the impact estimate is 0.10, which means that the bias is 0.10. This bias does not result in a false inference because the impact estimate is not statistically significant. As the sample size increases, both the impact estimate and the confidence interval shrink, corresponding to a decrease in bias and an increase in precision. However, the bias shrinks at a slower rate than the confidence interval such that the estimated (biased) impact eventually becomes statistically significant. Even though the bias has dropped from 0.10 to 0.03 by the time the sample

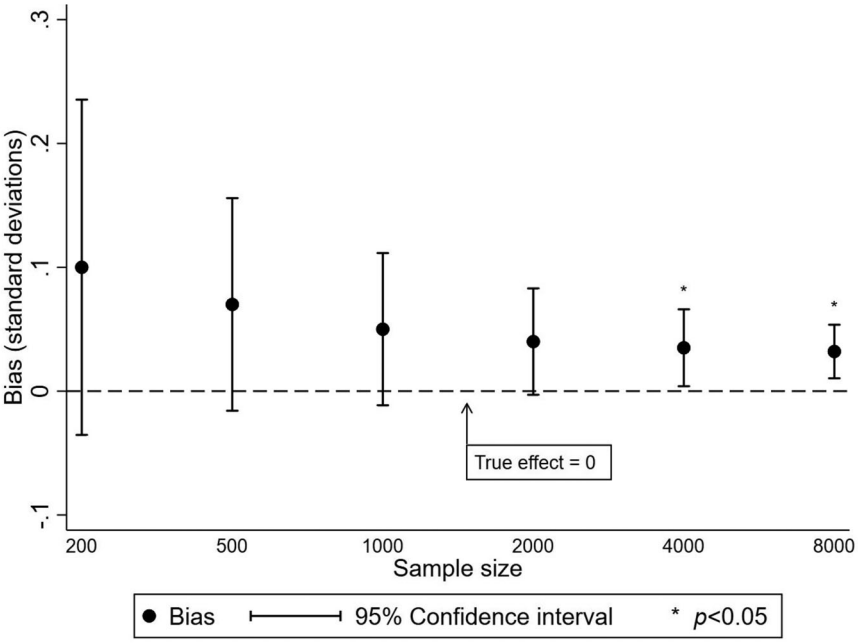

Figure 1. Illustrative example of asymdystopia. *Note*: The figure is a hypothetical example and not based on real data. The true effect is assumed to be zero; therefore, the bias is equivalent to the impact estimates.

size reaches 8,000, the study with the larger sample actually presents a greater risk for making false inferences—hence, asymdystopia.

In practice, asymdystopia could arise for a range of reasons in RCTs and RDDs.[6] Any bias that increases the absolute value of an impact estimate could lead to asymdystopia, assuming the standard error shrinks more rapidly than the bias as the sample size increases. Biases that attenuate the impact estimate would tend not to lead to asymdystopia because Type 1 errors would become *less* likely. Two common and important potential sources of upward bias in impact estimates are attrition in RCTs and functional form misspecification in RDDs. We focus on those two sources in this article to provide a concrete and detailed illustration of asymdystopia. However, it is important to note that asymdystopia is not limited to those two sources. Briefly, additional sources could include (but are not limited to):

Survey Response Bias

Studies can suffer from biased impact estimates if the intervention affects the measurement of an outcome but not the outcome itself. For example, Chen et al. (2020) conducted an experiment that demonstrated how an intervention might affect the measurement of self-reported social and emotional skills such as perseverance. Explaining the importance of such skills can increase students' reports of their own

[6]Asymdystopia can also arise in other causal impact designs, such as quasi-experiments (QEDs). For example, omitted variable bias can often lead to upwardly biased QED impact estimates, which likely do not diminish with sample size. A formal consideration of asymdystopia in QEDs is beyond the scope of this article, however.

skills by up to 0.11 standard deviations. The explanations occurred immediately before the students reported on their skills, suggesting that students' skills did not actually change. Interventions that target social and emotional skills often have this feature, potentially leading to upwardly biased impacts in practice. Because this form of bias likely does not decrease with sample size, asymdystopia could result.

Timing of Data Collection

Outcome data collected at systematically different times for treatment and control groups could result in bias. For example, if test scores are collected later for students in the treatment group, then their scores might be higher than the control group simply because test scores increase as students age. One such scenario is if tests are administered midsummer to students after they are treated with a monthlong academic enrichment program, whereas control students are tested at the end of the prior school year because they would otherwise be difficult to track down in the summer. If the logistical constraints that determine the timing of data collection are unrelated to sample size, then this potential source of bias could lead to asymdystopia.

Contamination

Contamination bias can occur if members of one study group are exposed to the condition of another study group. In the canonical example—such as when teachers share materials from a professional development treatment with control group teachers—there is a downward bias because the control group is exposed to an effective treatment, which dilutes the treatment-control contrast. While this bias unlikely shrinks with larger samples, it also unlikely leads to asymdystopia because Type 1 errors might actually *decrease* with sample size. However, contamination that biases impacts upward could lead to asymdystopia. For example, contamination between two treatment groups in a three-armed RCT could upwardly bias the impact of each treatment relative to control if the treatment combination is more effective than either treatment alone. As another example, impacts could be biased upward if there are negative spillovers from a partially-treated control group. One such scenario is an intervention that helps students identify and apply to colleges for which they are a good match, in order to boost college completion rates. If treatment students encourage their control group friends to apply to the same colleges and these colleges are on average a worse match than what their friends would have otherwise chosen, then the control group's college completion rate actually might end up being lower. The estimated impact of the intervention on target students would therefore be biased upward. These two examples of upward contamination bias would not likely decrease with sample size, and thus have the potential for asymdystopia.

Compromised Random Assignment

Staff conducting intake for an RCT might systematically assign individuals to the wrong experimental condition, which could lead to an upward bias. For example, intake staff might disregard the study's protocol and choose to assign more motivated students to the treatment group because they are most likely to benefit from an intervention.

If motivation is positively correlated with outcomes, then the resulting impact estimates would be biased upwards. This bias would not likely diminish with sample size, and thus could lead to asymdystopia.

Studies can simultaneously suffer from many of the biases considered above and other types of biases. In some cases, the biases may offset each other, but if they tend to be in the same upward direction, then even if each bias is small, they could compound and result in asymdystopia. However, as we discuss later, well-executed studies that carefully consider the possibility of these biases can mitigate the risk of asymdystopia.

How Problematic Is Attrition Bias in RCTs as Studies Are Powered to Detect Smaller Impacts?

Greenberg and Barnow (2014) identify sample attrition as potentially the most serious flaw that can lead to biased impact estimates in RCTs. Sample attrition occurs when individuals who were randomly assigned to treatment or control groups are missing outcome data. One key indicator of attrition bias is the attrition rate, both the overall rate for the study sample and the differential rates between treatment and control groups. Attrition bias is generally more concerning the larger the overall or differential attrition rates are. A second key indicator of attrition bias is how strongly attrition relates to outcomes and whether this relationship differs between treatment groups. For example, attrition bias would be high if outcome data were missing for the highest-achieving members of the control group and the lowest-achieving members of the treatment group. In general, the more strongly related these factors are, the more likely attrition bias is problematic. Unlike attrition rates, such relationships are at best incompletely observable, so some assumptions are needed to determine the risk of attrition bias for a given study. It is up to researchers to argue that the assumptions are plausible in their study's context.

To illustrate the problem posed by attrition bias in RCTs as studies are powered to detect smaller impacts, we first describe our model of attrition bias. Second, we examine how the tolerance level for overall and differential attrition changes as the target impact gets smaller. Third, we examine whether more favorable assumptions are needed about the relationship among attrition, outcomes, and treatment status as the target impact gets smaller. Finally, we examine the likely feasibility of executing studies powered to detect smaller impacts with attrition rates low enough to control bias at acceptable levels.

Summary of the WWC Attrition Model and Standard[7]

We base our analysis on an attrition model developed by the WWC (2013, 2014). Although attrition bias can be modeled in any number of ways, we feature this particular model because it cleanly illustrates the main issues and has been used to assess attrition bias in thousands of studies in education and other fields, making it a familiar and

[7]This summary draws heavily from the WWC's technical methods paper entitled *Assessing Attrition Bias* (https://ies.ed.gov/ncee/wwc/Document/243), which includes complete details of the attrition model.

relevant model for many readers.[8] To our knowledge, the intuition—if not the particularities—underlying our conclusions are model-agnostic and would likely apply to other models of attrition bias in RCTs, even more sophisticated ones.

The model begins by assuming that all study participants have an unobserved latent propensity to stay in the study. The lower this propensity, the more likely the study participant will attrite. This propensity, z, is assumed to be a normally distributed (0,1) random variable. If the total proportion of participants who stay in the study is denoted by P (and thus, the overall attrition rate is $1 - P$), and Φ is the standard normal cumulative distribution function, then participants will stay in the study if their z exceeds the threshold z^*, which is a deterministic function of P:

$$z > \Phi^{-1}(1 - P) = z^*. \tag{1}$$

The model further assumes y is the study outcome, also a normally distributed (0,1) random variable, and is related to z as follows:

$$\begin{aligned} y_t &= \alpha_t z_t + u_t \\ y_c &= \alpha_c z_c + u_c. \end{aligned} \tag{2}$$

Because this relationship may differ between treatment (t) and control (c) groups, there are two analogous equations subscripted by t and c; α is the correlation between z and y, and u is a normally distributed $\left(0,\ 1 - \alpha^2\right)$ random variable independent of z. If α is 1 or -1, then all of y can be explained by z, whereas if α is zero, then z has no influence on y. Thus, the closer α is to zero, the less attrition is related to study outcomes, and by extension, the less likely attrition would lead to biased impact estimates. The reverse is true as α gets closer to 1 or -1.

For simplicity, this model assumes that there are no impacts on mean outcomes in the study sample. Because there are no true impacts, an unbiased estimator should find no differences in expectation between treatment group outcomes and control group outcomes. Thus, attrition bias (B) is simply the expected difference between treatment group outcomes (y_t) and control group outcomes (y_c), which can be expressed using the following analytic formula, based on the properties of truncated normal distributions (ϕ is the standard normal probability density function):

$$\begin{aligned} B &= E\left(y_t | z_t > z_t^*\right) - E\left(y_c | z_c > z_c^*\right) = \alpha_t E\left(z_t | z_t > z_t^*\right) - \alpha_c E\left(z_c | z_c > z_c^*\right) \\ &= \frac{\alpha_t \times \phi\left(\Phi^{-1}(1 - P_t)\right)}{P_t} - \frac{\alpha_c \times \phi\left(\Phi^{-1}(1 - P_c)\right)}{P_c}. \end{aligned} \tag{3}$$

This result shows that attrition bias is driven by two main factors: the fraction of non-attriters (P_t and P_c), and the strength of the relationship between attrition and outcomes (α_t and α_c). Moreover, the *differences* in these factors across treatment and control groups are important to consider. For example, if $P_t = P_c$ and $\alpha_t = \alpha_c$, there will be no attrition bias, even if a large proportion of the sample leaves and even if attrition is strongly related to outcomes. This result arises because the same types and fractions of participants drop out of both treatment and control groups. This uniformity preserves the equivalence of the remaining participants across both groups, leading to unbiased

[8]The U.S. Department of Health and Human Services has also used this model. See, for example, the Home Visiting Evidence of Effectiveness Review (http://homvee.acf.hhs.gov) and the Teen Pregnancy Prevention Evidence Review (http://tppevidencereview.aspe.hhs.gov).

impact estimates. However, if either $P_t \neq P_c$ or $\alpha_t \neq \alpha_c$, then attrition bias will generally be present. If outcomes and the propensity to respond are not normally distributed but are correlated with each other, then attrition bias will still arise as long as $E(y_t|z_t>z_t^*) - E(y_c|z_c > z_c^*) \neq 0$.

The analytic formula for attrition bias in Equation (3) allows us to precisely map out how much attrition bias exists for different combinations of attrition rates and α's. The WWC uses two sets of assumptions for α. The conservative assumption sets $\alpha_t = 0.45$ and $\alpha_c = 0.39$. The optimistic assumption sets $\alpha_t = 0.27$ and $\alpha_c = 0.22$. The conservative and optimistic assumptions differ in two ways: (1) the degree to which study participants with outcome data differ from those without outcome data (i.e., the size of α_t and α_c) and (2) the extent to which that relationship is itself related to treatment status (i.e., how large the *difference* between α_t and α_c is). The optimistic assumption has a lower overall α_t and α_c, and a smaller difference between α_t and α_c. These assumptions imply that attrition is less related to the outcome and less related to treatment status, which suggests that all else equal, attrition bias would be less problematic.

It is not possible to estimate α_t and α_c directly. The WWC did, however, validate these parameter values based on empirical correlations between attrition and *baseline* measures of outcome variables, used as a proxy for the correlation between attrition and *follow-up* measures of those outcome variables. These correlations came from large-scale experimental evaluations of seven interventions (six curricular interventions and one teacher certification intervention) covering multiple grades and outcomes. They found that the observed correlations were generally most consistent with the optimistic assumption, but they retained the conservative assumption for special cases in which the treatment might plausibly have significant impacts on attrition.

For each of the two assumptions for α, it is possible to use Equation (3) to calculate the bias for various combinations of overall and differential attrition rates. More formally, the overall attrition rate is the proportion of randomized study participants who lack data on the evaluation's outcomes (equivalent to $1 - P$ in Equation (3)). The differential attrition rate is the difference between the treatment and control groups in the proportion of randomized study participants who lack data on the evaluation's outcomes (equivalent to $P_t - P_c$ in Equation (3)). If the goal is to keep attrition bias within a certain maximum acceptable level, this exercise will reveal the acceptable combinations of overall and differential attrition rates. This method is exactly how the WWC derived its attrition standard.

The attrition standard aims to keep attrition bias to no more than 20% of the impact. Because the WWC defines a substantively important impact as 0.25 standard deviations for the purposes of attrition, the maximum acceptable level of attrition bias is 0.05 standard deviations. By keeping attrition bias at this level, the Type 1 error rate is controlled at about 8% in studies that conduct hypothesis testing at the 5% significance level and that are powered to detect an impact of 0.25 standard deviations (with 80% power). In other words, the real Type 1 error rate is 8% compared to the nominal rate of 5%.

Panel (a) of Figure 2 highlights the resulting bounds on overall and differential attrition rates. The black region shows combinations of overall and differential attrition rates that yield attrition bias less than or equal to 0.05 standard deviations under the

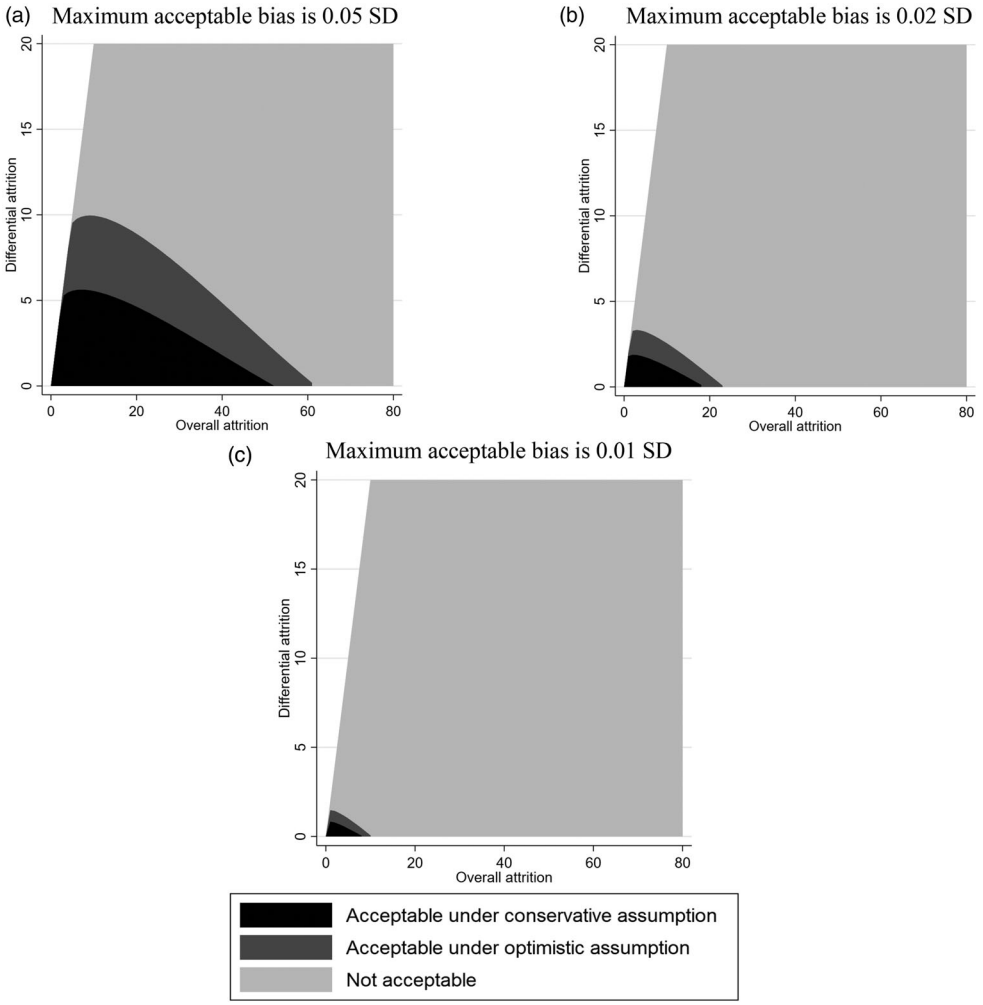

Figure 2. Attrition bounds. *Note*: Authors' calculations. The black and dark gray regions represent values of overall and differential attrition rates that are acceptable under the conservative and optimistic assumptions, respectively. The light gray region indicates values of overall and differential attrition rates that are not acceptable.

conservative assumption. The dark gray region shows combinations that yield acceptable bias under the optimistic assumption. The light gray region shows combinations that yield unacceptable bias under both sets of assumptions. Thus, to meet the WWC attrition standard, researchers have tried to keep overall and differential attrition rates within the black or dark gray regions.

Acceptable Attrition Rates for Studies Powered to Detect Small Impacts

Staying within the black and dark gray regions in Figure 2 helps ensure that attrition bias is no larger than the maximum acceptable bias of 0.05 standard deviations. However, as studies are powered to detect impacts smaller than 0.25 standard

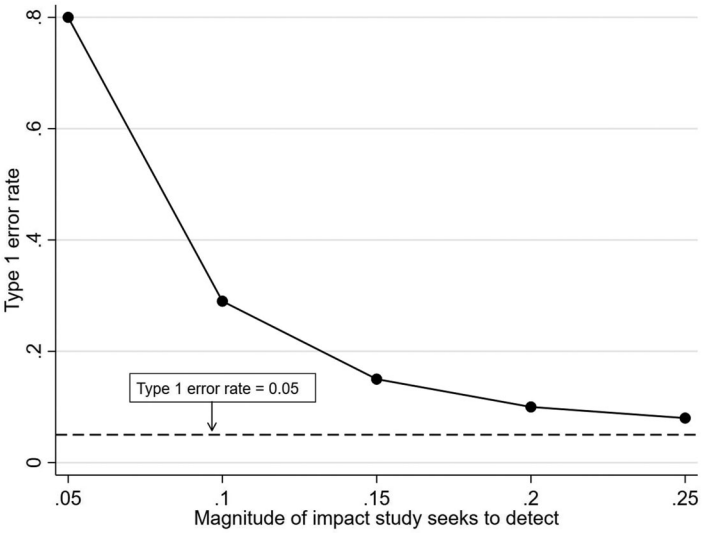

Figure 3. The Type 1 error rate increases as studies are powered to detect smaller effects if attrition bias is held constant at 0.05 standard deviations. *Note*: Authors' calculations. These calculations assume an RCT designed to detect a substantively important impact with 80% power at a significance level of 5%. The figure shows that as studies are powered to detect smaller effects, the Type 1 error rate increases if attrition bias is held constant at 0.05 standard deviations.

deviations, the maximum acceptable bias also needs to be reduced accordingly to ensure that attrition bias accounts for no more than 20% of the impact and that the Type 1 error rate is controlled at an acceptable level. This means that if a study is powered to detect an impact of 0.10 standard deviations, attrition bias should be limited to 0.02 standard deviations. Similarly, if a study is powered to detect an impact of 0.05 standard deviations, attrition bias should be limited to 0.01 standard deviations. If the maximum acceptable bias is not reduced, then attrition bias could account for most or all of the estimated impact, which would lead to a much higher Type 1 error rate (Figure 3), even if actual attrition levels fall within the black or dark gray regions in panel (a) of Figure 2.

To show how reducing the maximum acceptable bias from 0.05 to 0.02 or 0.01 affects attrition levels, we re-shade the black, dark gray, and light gray regions of panel (a) of Figure 2. In panel (b) of Figure 2, we shade the areas that based on Equation (3), yield bias of no more than 0.02 standard deviations (instead of 0.05). In panel (c) of Figure 2, we shade the areas that yield bias of no more than 0.01 standard deviations. The results in panels (b) and (c) of Figure 2 show that substantially tighter attrition bounds are needed. For example, assuming (1) the WWC's optimistic parameters, (2) no differences in attrition rates between treatment and control groups, and (3) a maximum acceptable bias of 0.05, the highest acceptable overall attrition rate is about 60% (panel (a) of Figure 2). All else equal, if the maximum acceptable bias is 0.02 instead of 0.05, then the analogous highest acceptable overall attrition rate drops from 60% to about 20% (panel (b) of Figure 2). If the maximum acceptable bias is 0.01, then the highest acceptable overall attrition rate drops to about 10% (panel (c) of Figure 2).

Table 1. Highest acceptable differential attrition rate.

Highest acceptable bias (standard deviations)	Half of highest acceptable overall attrition rate (%)	Highest acceptable differential attrition rate (percentage points)
0.05	30	6
0.02	10	2
0.01	5	1

Note: Authors' calculations. Highest acceptable overall attrition rate is the highest level of attrition at which bias is below the highest acceptable level when there is zero differential attrition.

The highest acceptable *differential* attrition rate is also substantially smaller when limiting the maximum acceptable bias to 0.02 or 0.01. For example, in Table 1 we calculate the highest acceptable differential attrition rate, assuming that the overall attrition rate is half of the maximum acceptable overall rates presented in the previous paragraph's example. Under the WWC's optimistic assumptions—an overall attrition rate of 30%, and a maximum acceptable bias of 0.05—the highest acceptable differential attrition rate is about 6% points (e.g., where the treatment group attrition rate is 33% and the control group attrition rate is 27%). If the maximum acceptable bias is 0.02 instead of 0.05, then the highest acceptable differential attrition rate is 2% points, rather than 6% points. If the maximum acceptable bias is 0.01, then the highest acceptable differential attrition rate is about 1% point.

Accounting for the Study Context When Determining Acceptable Attrition Rates

The previous section's results show that substantially lower levels of overall and differential attrition are needed to contain bias in studies powered to detect small impacts, given the WWC optimistic parameter assumptions. However, more favorable assumptions may be justifiable in some studies. If so, bias could still be contained to an acceptable level in these studies *even if the overall and differential attrition levels are in the standard WWC ranges* (panel (a) of Figure 2).

In this section, we examine just how much more favorable these assumptions would need to be for the standard WWC attrition bounds to be appropriate for studies powered to detect small impacts. To do so, we use Equation (3) to compute which values of α_t and α_c will contain bias to the lower levels needed (i.e., 0.01 or 0.02 standard deviations, instead of the usual 0.05) in studies powered to detect small impacts, assuming that attrition levels fall within the typical bounds in panel (a) of Figure 2.[9] Table 2 reports the results for one set of attrition rates, but the basic conclusion that α_t and α_c would need to be more favorable holds more generally across all combinations of attrition rates. Recall that smaller overall α's and smaller differences between α_t and α_c are more favorable because they imply that attrition is less related to both outcomes and treatment status, and therefore less likely to bias the estimated impacts. The results clearly show that as the maximum acceptable attrition bias falls for studies powered to detect small impacts, the model assumptions need to become more favorable for any given level of overall and differential attrition.

[9]For any observed overall and differential attrition rates, there are many values of α_t and α_c that would yield a given level of bias (see Equation (3)). To calculate a unique pair of model parameters for each given level of bias, we assume that $\alpha_t = r\alpha_c$, where r is a constant equal to the ratio of α_t to α_c implicit in the WWC parameters (0.27/0.22). This approach allows us to uniquely characterize how optimistic the study parameters would need to be to contain bias.

Table 2. Assumptions needed to apply current attrition bounds with lower bias.

	Acceptable attrition		Attrition model parameter assumptions	
Maximum acceptable bias	Overall	Differential	α_t	α_c
0.05	30	6	0.27	0.22
0.02	30	6	0.12	0.10
0.01	30	6	0.06	0.05

Note: Authors' calculations using attrition model described in Equation (3). The first row corresponds to the existing WWC optimistic attrition standard, which seeks to contain bias to 0.05 standard deviations. The second and third rows show how attrition model parameter assumptions would need to change to limit bias to 0.02 and 0.01 standard deviations at the same levels of overall and differential attrition. Values of α_t and α_c are correlations, the attrition rates are percentage points, and the maximum acceptable bias is standard deviation units.

Just how much more favorable are these assumptions? As noted earlier, the WWC calculated the correlations between attrition and baseline measures in a number of education studies and then used these correlations as a proxy for the correlation between attrition and outcome measures. Across seven interventions, they found that the correlation between baseline measures and attrition ranged from 0.01 to 0.28 for treatment groups and from 0.06 to 0.26 for control groups. Moreover, the treatment–control difference in correlations ranged from 0.01 to 0.10. Using these benchmarks, we see that the required assumptions calculated in Table 2 for lower levels of attrition bias (0.01 and 0.02) are within the empirically observed ranges, although they are at the more optimistic end of that range.

To gain an even better understanding of how much more optimistic these assumptions are, we simulated outcome and attrition data using the attrition rates and values of α_t and α_c shown in Table 2. These data were generated using the formulas in Equations (1) and (2), which means that the outcomes for the full sample (including both attriters and non-attriters) follow the standard normal distribution (mean zero, variance one).[10] Table 3 shows the results for three different scenarios. For scenario 1, we generated data with optimistic WWC values for α_t and α_c, and attrition rates for the treatment and control groups that yield bias of 0.05 standard deviations. We report the mean of the outcome variable for the attrited and non-attrited samples in the treatment and control groups, as well as the difference in these means. Scenarios 2 and 3 hold the attrition rates constant but change the values of α_t and α_c to yield biases of 0.02 and 0.01 standard deviations.

Table 3 shows that to apply the existing WWC attrition bounds for lower levels of acceptable bias, we must effectively assume that the participants who leave a study's sample are increasingly similar to those who stay and that the participants who leave the treatment group are increasingly similar to those who leave the control group. First, there is a much smaller difference in outcomes between participants who leave the study and those who stay. Under the WWC optimistic assumptions (scenario 1), follow-up test scores of participants who leave the study are about 0.37 to 0.45 standard deviations lower than those of participants who stay. But under the assumptions needed to limit bias to 0.02 or 0.01 standard deviations (scenarios 2 and 3), this gap must fall to as little as 0.08 to 0.10 standard deviations. Second, Table 3 shows a smaller difference between the attrited samples for the treatment and control groups (meaning that the intervention

[10]Note that it does not matter which attrition rates and values of α correspond to the treatment or control groups—switching all treatment and control labels would yield the same conclusions.

Table 3. Outcomes of attrited and non-attrited samples generated under varying assumptions.

		Outcomes of attriters and non-attriters		
α	Attrition rate	Mean of attriters	Mean of non-attriters	Difference in means (mean of non-attriters—mean of attriters)
Scenario 1: Attrition bias of 0.05 under WWC optimistic parameter assumptions				
$\alpha_t = 0.27$	33	−0.30	0.15	0.45
$\alpha_c = 0.22$	27	−0.27	0.10	0.37
Scenario 2: Parameter assumptions that yield attrition bias of 0.02 under scenario 1 attrition rates				
$\alpha_t = 0.12$	33	−0.13	0.06	0.19
$\alpha_c = 0.10$	27	−0.12	0.05	0.17
Scenario 3: Parameter assumptions that yield attrition bias of 0.01 under scenario 1 attrition rates				
$\alpha_t = 0.06$	33	−0.07	0.03	0.10
$\alpha_c = 0.05$	27	−0.06	0.02	0.08

Note: Authors' calculations using attrition model described in Equation (3). The first row for each scenario is the treatment group, the second row is the control group. Values of α_t and α_c are correlations, the attrition rates are percentage points, and the descriptive statistics are standard deviation units.

had a smaller differential impact on the types of participants who left the treatment group versus the control group). This difference was already modest under WWC assumptions (0.03 standard deviations under scenario 1, the difference between −0.30 and −0.27), but it becomes even smaller (0.01 standard deviations) under the more favorable scenarios needed for studies powered to detect smaller impacts (scenarios 2 and 3).

Evidence on the Feasibility of Attaining Acceptable Attrition Rates

When researchers design studies to detect smaller impacts and still want to ensure that attrition bias accounts for no more than 20% of their smallest detectable impact, they need to consider whether they can realistically achieve lower attrition rates. To investigate whether lower attrition is feasible in practice, we used the study review database from the WWC to examine how often past studies achieved overall and differential attrition rates consistent with limiting bias to no more than 0.02 or 0.01 standard deviations (corresponding to study MDEs of 0.10 or 0.05 standard deviations).

From the WWC database, we focused on RCTs that received a rating of Meets WWC Standards Without Reservations because these represent well-executed studies that provide a natural benchmark for considering the feasibility of achieving lower levels of attrition.[11] We supplemented the WWC downloadable database with additional information on sample sizes and attrition rates from the WWC master review guides. We focused our analysis on each study's main impact estimates, which were the basis for the WWC's rating. Information on supplementary analyses, such as the impacts on subgroups, were excluded. For each study, we calculated the MDE using the *p*-value, effect size, and analytical sample size.[12]

[11]We exclude quick reviews because the review protocol differs from other types of reviews. The database is available at https://ies.ed.gov/ncee/wwc/StudyFindings.

[12]Typically the MDE is expressed as a function of the standard error. However, the WWC does not record standard errors, so we infer the standard error from the combination of the impact estimate, *p*-value, and analytical sample size. For each study, we calculated the MDE using the following formula: $MDE = \left|T^{-1}\left(N-1, 1-\frac{\alpha}{2}\right) + T^{-1}(N-1, \beta)\right| \times \left|ES/T^{-1}\left(N-1, \frac{p}{2}\right)\right|$, where T^{-1} is the inverse *t*-distribution, α is the significance level (assumed to be 0.05), β is the power (assumed to be 0.80), *ES* is the effect size, *p* is the *p*-value, *N* is the analytical sample size for the unit of randomization, and the vertical bars indicate the absolute value.

Table 4. The percentage of past studies with acceptable attrition under three different maximum acceptable bias thresholds and three attrition model parameters assumptions.

Attrition model parameters		Percentage of past studies with acceptable attrition under three maximum acceptable bias thresholds		
α_t	α_c	0.01	0.02	0.05
0.06	0.05	92	100	100
0.12	0.10	61	92	100
0.27	0.22	33	57	95

Note: Authors' calculations using WWC database and supplemental attrition data from WWC master review guides.

Studies with low MDEs (which we define to mean less than 0.15 standard deviations) represent approximately 20% of all studies (170 out of 869 studies). Under the WWC's optimistic parameter assumptions, over half of studies with low MDEs have attrition rates low enough to keep bias below 0.02 standard deviations, and one-third have attrition rates low enough to keep bias below 0.01 standard deviations (Table 4, row 3). In study contexts where even more optimistic assumptions are appropriate, these percentages can be much higher. With a bias threshold of 0.02, 92% of studies have acceptable attrition under the more optimistic parameters considered earlier, $\alpha_t = 0.12$ and $\alpha_c = 0.10$ (Table 4, row 2, column 2). With a bias threshold of 0.01, 92% of studies have acceptable attrition under the most optimistic parameters considered earlier, $\alpha_t = 0.06$ and $\alpha_c = 0.05$ (Table 4, row 1, column 1). These findings suggest that in many cases researchers can feasibly attain attrition levels that are low enough to limit biases to lower levels, especially if more optimistic parameter assumptions are warranted.

Researchers should carefully consider whether their study context warrants more optimistic parameter assumptions. If more optimistic assumptions are made when they are unwarranted, the result could be a low-quality study with misleading findings. Recall that attrition is particularly problematic when students with missing data in the treatment group are fundamentally different from students with missing data in the control group. There are several scenarios where this is possible, including the following:

1. **High-ability students assigned to a control group in a charter school evaluation move to a private school.** In a study of charter schools that relies on administrative data from school districts for test score outcomes, some parents whose children are not accepted into the charter school through a randomized lottery might look for opportunities to move their children into a private school outside of the study. This reaction to the lottery could result in the best students leaving the control group but not the treatment group, creating the illusion of a positive impact.
2. **Teachers in the treatment group discourage low-ability students from taking an achievement test.** In a study of financial incentives for teachers whose students show the highest performance gains, teachers in the treatment group might have an incentive to discourage low-ability students from taking the test used to measure the teacher's performance.
3. **A dropout prevention program keeps lower-ability students in school in the treatment group, resulting in biased impacts on academic achievement outcomes.** By design, a dropout prevention program is intended to affect whether

students remain in school, which in turn can affect attrition because dropouts often have missing data. If the program is successful, then the treatment group may include students who would have dropped out had they been in the control group. This phenomenon could result in a different mix of students taking achievement tests in the treatment and control groups.

The bottom line is there are compelling reasons for researchers to continue conducting studies that are powered to detect small impacts, but researchers should be more attuned to the threat of attrition bias in these studies. To adequately contain potential bias and the risk of making false inferences, researchers should be prepared to invest additional resources to keep attrition at levels below what is typical for many past studies that have been powered to detect small impacts. Researchers might also (carefully) consider whether more optimistic assumptions about the attrition process are warranted in their study than what has been typical in prior studies of education interventions. More optimistic assumptions allow for attrition levels that are in the range of what past studies have experienced. Researchers can use formal attrition models such as the one developed by the WWC as a tool for assessing both the level of attrition that is acceptable for a given set of assumptions about the attrition process and how optimistic these assumptions need to be for a given level of attrition.

Is Functional Form Misspecification Bias More Problematic in RDDs That Are Powered to Detect Small Impacts?

Under an RDD, a cutoff on a continuous assignment variable is used to determine who is offered the opportunity to participate in a program. If the program has an impact, we would expect to see an abrupt change—a "discontinuity"—in the outcome at the cutoff. For example, because of funding constraints, a school district might only provide free after-school math tutoring to students scoring below a cutoff on a pretest, creating the opportunity to estimate the impact of math tutoring using an RDD. Students with scores below the cutoff would be in the treatment group; students above the cutoff would be in the comparison group. A valid estimate of the impact of math tutoring could then be obtained by comparing the outcomes of students below and above the cutoff, after adjusting for students' pretest scores.

Unlike an RCT, the validity of an RDD hinges on statistical modeling, specifically modeling of the relationship between the outcome and the assignment variable. For example, if the true relationship between the outcome and the assignment variable is not linear, then fitting a linear regression line to all of the data on either side of the cutoff might result in a biased impact estimate. Mainstream RDD methods typically try to address functional form misspecification bias by selecting a bandwidth (or narrow window) around the treatment–comparison cutoff and estimating a linear or quadratic regression within the bandwidth (Calonico et al., 2014; Gelman & Imbens, 2019; Imbens & Kalyanaraman, 2012). Generally, smaller bandwidths yield less functional form misspecification bias because linear or quadratic approximations become more appropriate as bandwidths get smaller. However, smaller bandwidths also include fewer data points, thus adversely affecting the precision of the estimate.

To manage the tradeoff between bias and precision, mainstream algorithms typically attempt to choose a bandwidth that minimizes the mean squared error (e.g., Calonico et al., 2014; Imbens & Kalyanaraman, 2012). The mean squared error is the square of bias plus the variance of the impact estimate. Because the objective is to minimize the *sum* of these two components, there is no guarantee that each component will decrease in equal proportion. If, as a study becomes larger, the standard error of the impact estimate shrinks more quickly than the functional form misspecification bias (i.e., precision increases much faster than bias shrinks), then Type 1 errors could become more common. In other words, even though these mainstream algorithms have desirable properties—namely, they are data-driven and select bandwidths that yield *asymptotically* unbiased impact estimates—a naïve application of them could lead to asymdystopia, where studies with larger (but still finite) sample sizes are actually at greater risk of making false inferences.

For example, consider two studies of different sizes in which an RDD is used to test an education intervention that truly has no impact on student achievement. In one RDD study, there is a sample of 500 students, and the researcher estimates an impact of 0.06 standard deviations with a standard error of 0.04, which is not statistically significant at conventional levels. In the second RDD study, there is a larger sample of 5,000 students, and the researcher estimates an impact of 0.04 with a standard error of 0.02, which is statistically significant at conventional levels. In this example, the larger study is "better" in the sense that the bias in the impact estimate is smaller (0.04 versus 0.06 relative to a true null impact). On the other hand, the larger study is also "worse" because it leads to a Type 1 error.

Concerns about accurate confidence interval coverage for RDD estimators are not new. For example, Calonico et al. (2014) suggest a technique for adjusting the impact estimates and standard errors derived from mainstream bandwidth selection algorithms that control Type 1 errors at the desired rate. In addition, there is a rapidly growing literature on state-of-the-art RDD estimators, many of which are designed to automatically address coverage concerns and appear to be even more optimal than mainstream methods.[13] Some of these recent methods employ frameworks that differ from mainstream methods in fundamental ways, such as using optimality criteria other than MSE-minimization (Armstrong & Kolesár, 2020; Calonico et al., 2020; Sales & Hansen, 2019). While investigating these state-of-the-art methods is beyond the scope of this article, it is important for readers to be aware of more recent methods that are also relevant to the education contexts and issues we explore.

To investigate the potential for asymdystopia in RDDs, we examine how misspecification bias changes as the size of an RDD study increases under two mainstream RDD methods (Calonico et al., 2014; Imbens & Kalyanaraman, 2012). We focus on these nonparametric local regression estimators that use data-driven bandwidth selectors, not because they are necessarily the best methods, but because they are commonly accepted and used by applied researchers (Pei et al., 2020; WWC, 2020). More specifically, we use Monte Carlo simulations to assess bias under varying assumptions regarding the

[13]A non-exhaustive list of relevant papers includes: Armstrong and Kolesár (2020); Bartalotti (2019); Bartalotti et al. (2017); Branson et al. (2019); Calonico et al. (2020); Cattaneo et al. (2015, 2017); He and Bartalotti (2020); Imbens and Wager (2019); Noack and Rothe (2020); Sales and Hansen (2019).

true relationship between the outcome and assignment variable. We examine whether statistical power increases with sample size, as well as how the magnitude of functional form misspecification bias changes and how the Type 1 error rate changes.

The purpose of this exercise is to better understand, based on real-world education contexts, (1) the extent to which misspecification bias increases the risk of making false inferences in RDD studies that are powered to detect small impacts, and (2) whether this potential problem is indeed mitigated by the technique for adjusting impact estimates and standard errors that Calonico et al. (2014) suggest. The purpose of this exercise is *not* to suggest that all else equal, researchers should prefer smaller studies to larger ones, as there are many cases where a larger sample size is needed to detect impacts of a meaningful magnitude.

Methodological Approach

Our methodological approach is to use Monte Carlo simulations, where we generate data through a known but random process and then estimate RDD impacts, standard errors, and *p*-values using two mainstream approaches. After repeating this process many times, we assess how the different approaches perform under a variety of realistic conditions that education researchers may face when conducting evaluations using an RDD.

In our Monte Carlo simulations, we generate data using seven different data generating processes (DGPs). To make the simulation findings more relevant to researchers, the DGPs are based on data from previous education studies that included math and reading post-tests and pretests. Each DGP consists of a fifth-order polynomial equation that describes the relationship between the assignment variable (pretest) and the outcome (post-test). The cutoff used in each case is the median value of the pretest. Each DGP also describes the distribution of the assignment variable, including whether and how individuals are clustered within unique values of the assignment variable. Finally, each DGP specifies what proportion of the variance of the outcome is due to the assignment variable versus unobserved random factors. Details regarding the DGPs are reported in the appendix. The specific steps of our simulation procedure are as follows:

1. Generate three data sets for each of the seven DGPs specified in the appendix. One data set has 1,000 observations, the second has 10,000 observations, and the third has 100,000 observations.
2. Estimate RDD impacts on each of the simulated data sets using two different bandwidth selection algorithms and two different approaches to calculating standard errors. The two bandwidth algorithms are those suggested by Imbens and Kalyanaraman (2012) and Calonico et al. (2014).[14] The standard error

[14]In some cases, these algorithms select bandwidths that are so narrow there are not enough data to calculate an impact and/or a standard error. In those cases, we automatically expand the bandwidth until we can calculate an impact and standard error. Of the seven DGPs (see Figures A3 and A4), three had cases where the bandwidth had to be expanded. For the DGPs represented in Figure A3 panes (b) and (c), the bandwidth had to be expanded in 78% of Monte Carlo replications when the CCT bandwidth selection algorithm was used with a sample size of 100,000. For the DGP represented in Figure A4, pane (d), the bandwidth had to be expanded in up to 1% of replications when the IK bandwidth was used (regardless of sample size). When the CCT algorithm was used, the bandwidth had to be

Table 5. Summary of findings from simulations based on data from education studies.

	Bandwidth selection algorithm and sample size					
	Imbens and Kalyanaraman (2012)			Calonico et al. (2014)		
Standard errors	1,000	10,000	100,000	1,000	10,000	100,000
Magnitude of the functional form misspecification bias (in standard deviations)						
Conventional						
Average across outcomes	0.012	0.010	0.007	0.010	0.007	0.006
Range across outcomes	0.000–0.035	0.001–0.030	0.000–0.030	0.001–0.033	0.001–0.030	0.000–0.030
Robust						
Average across outcomes	0.003	0.003	0.003	0.004	0.003	0.002
Range across outcomes	0.000–0.007	0.000–0.008	0.000–0.008	0.001–0.009	0.001–0.008	0.000–0.008
Minimum detectable effect (in standard deviations)						
Conventional						
Average across outcomes	0.469	0.153	0.053	0.543	0.173	0.061
Range across outcomes	0.435–0.506	0.137–0.173	0.043–0.067	0.484–0.679	0.154–0.221	0.049–0.077
Robust						
Average across outcomes	0.866[a]	0.256[a]	0.076	0.683	0.220	0.081
Range across outcomes	0.767–1.010	0.237–0.282	0.071–0.088	0.590–0.827	0.188–0.280	0.060–0.101
Type 1 error rate (the target is 0.05)						
Conventional						
Average across outcomes	0.060	0.067	0.114	0.059	0.059	0.105
Range across outcomes	0.058–0.062	0.053–0.096	0.048–0.398	0.051–0.064	0.054–0.077	0.051–0.398
Robust						
Average across outcomes	0.051	0.049	0.052	0.052	0.050	0.053
Range across outcomes	0.038–0.056	0.046–0.053	0.048–0.054	0.039–0.057	0.047–0.054	0.048–0.056

Note: Authors' calculations. The findings reported in this table are averaged across seven Monte Carlo simulations with 10,000 replications corresponding to seven data generating processes (DGPs). The robust estimation approach included bias-corrected impact estimates and standard errors inflated to control the coverage error rate, as suggested by Calonico et al. (2014).

[a]Three out of the 10,000 Monte Carlo replications for one of the seven DGPs yielded extremely large standard errors that severely skewed these values. Those extreme outliers were removed from the calculation of this average minimum detectable effect.

estimation approaches are (1) a "conventional" approach that ignores finite sample bias and (2) an approach that uses Calonico et al.'s method for calculating bias-corrected impact estimates and robust standard errors.

3. Repeat steps 1 and 2 10,000 times, recording impacts, standard errors, *p*-values, and bandwidth estimates.

With thousands of simulated impact estimates, we can look at summary statistics of how the estimates perform under varying conditions. We report three sets of findings:

1. **The average functional form misspecification bias across Monte Carlo replications.** Because data are generated under the null hypothesis of no "true" impact, the mean bias is equal to the mean estimated RDD impact.
2. **The average MDE across Monte Carlo replications assuming 80% power.** Using the estimated standard error of the RDD impact estimate for each replication, we calculate the smallest impact that would, with high probability, be statistically significant at the 5% level—this is just 2.8 times the estimated standard

expanded in 70% of replications with a sample size of 1,000 and over 99% of replications with a sample size of 10,000 or 100,000.

error for a two-sided hypothesis test (Bloom, 1995). These MDEs indicate the precision of the simulated RDD studies.

3. **The average Type 1 error rate across Monte Carlo replications.** This is the proportion of statistically significant impact estimates (i.e., where the *p*-value of the impact is less than 0.05). In an empirical approach with appropriate inference, this false inference rate should be 0.05, as the model assumes no "true" impact.

Simulation Findings

Our simulation findings confirm that with conventional estimation, Type 1 error rates increase as studies are powered to detect smaller impacts, but that the robust estimation approach that Calonico et al. (2014) recommend does in fact solve this problem. In Table 5, we report a summary of findings averaged across the seven DGPs for both conventional and robust estimation, and for bandwidths selected using either Imbens and Kalyanaraman (2012) or Calonico et al. (2014) algorithms. Note that Calonico et al.'s bandwidth selection algorithm is distinct from the robust procedures that the authors also recommend for estimating impacts and standard errors.

Conventional Estimation Findings

The MDE shrinks (and precision increases) as the sample size increases, as expected. Bias also shrinks, but at a slower rate than the MDE, resulting in an increasing rate of Type 1 errors. With Calonico et al. (2014) bandwidth selection algorithm and a sample size of 1,000, bias is 0.01 standard deviations, the MDE is 0.543 standard deviations, and the Type 1 error rate is 0.059. With a sample size of 100,000, bias shrinks to 0.006 standard deviations, the MDE shrinks to 0.061 standard deviations, and the Type 1 error rate increases to 0.105. The pattern of findings using Imbens and Kalyanaraman (2012) bandwidth selection algorithm is similar.

Robust Estimation Findings

Both bias and the MDE shrink as the sample size increases, as was the case with the conventional estimation findings. In this case, however, the shrinkage rates are similar, and the Type 1 error rate is not adversely affected. With Calonico et al. (2014) bandwidth selection algorithm and a sample size of 1,000, bias is 0.004 standard deviations, the MDE is 0.683 standard deviations, and the Type 1 error rate is 0.052. With a sample size of 100,000, bias shrinks to 0.002 standard deviations, the MDE shrinks to 0.081 standard deviations, and the Type 1 error rate is 0.053. The pattern of findings using Imbens and Kalyanaraman (2012) bandwidth selection algorithm is similar.

Discussion

In this article, we have demonstrated that although it is often desirable to conduct a study capable of detecting small impacts, researchers should be aware of the greater risk of false inference due to small biases. This concern applies to two of the strongest possible evaluation designs—RCTs and RDDs. We have shown that as studies are powered

to detect smaller impacts, some types of bias that previously might have been negligible can become significant threats to the credibility of a study's findings. This is because although statistical power generally increases with sample size, some sources of bias *do not decrease* with sample size (in the case of attrition bias in RCTs) or do not always decrease as quickly as power increases (in the case of functional form misspecification bias in RDDs). Thus, the *relative* threat of these biases can become larger in studies that are powered to detect smaller impacts.

These findings should not be interpreted as promoting smaller studies, but rather as encouraging appropriate care when designing larger studies. All else being equal, more statistical power is always better. For example, in a study that seeks to detect an impact of 0.10 standard deviations, it would be better to have 80% power than 60% power. Fortunately, with proper awareness and action, researchers can mitigate these threats. Below we discuss strategies that researchers can consider using. We focus on the two particular sources of bias that this article considered in depth (attrition in RCTs and functional form misspecification in RDDs). However, a similar thought process can be applied to any other source of bias, and our final section discusses mitigation strategies in more general terms.

Strategies to Address Small Biases Due to Attrition in RCTs

In the case of attrition bias in RCTs, we suggest three strategies. First, researchers can mitigate bias by expending more resources to achieve higher response rates for the collection of outcome data. However, even with substantially greater study resources, it might not always be possible to reduce attrition to the extent necessary because there may be diminishing marginal returns for each additional dollar invested in reducing attrition.

Second, attrition bias could be partially mitigated in some studies by statistically adjusting for observed differences in baseline characteristics between those who do and do not attrite, and how that difference varies between the treatment and control groups. Puma et al. (2009) examine several approaches to account for missing outcome data, including multiple imputation, regression adjustment, and nonresponse weights. These analytic adjustments will be most effective when researchers have access to baseline data that are correlated with both outcomes and attrition.[15]

Third, in some contexts researchers might be able to make more optimistic assumptions regarding the negative consequences of attrition. Attrition models, such as the one developed by the WWC and used in federal evidence reviews, can provide a framework for incorporating these assumptions into an assessment of acceptable levels of attrition. However, more optimistic assumptions should only be made when appropriate for the study context. Making more optimistic assumptions when they are unwarranted may worsen the problem by increasing the risk of misleading findings.

When considering the second strategy, researchers could conduct an empirical check to examine how correlated attrition is with baseline measures of the outcome variables, which serve as a proxy for actual outcomes. Researchers could also examine the

[15]The WWC attrition model does not directly incorporate covariates. However, the benefits of adjusting for covariates can be reflected in the model by making more optimistic assumptions regarding the negative consequences of attrition.

differences in baseline characteristics between attriters and non-attriters, as well as differences in baseline characteristics between attriters from the treatment group and attriters from the control group. However, we strongly caution against taking these empirical checks as absolute truth, as they may not be precisely estimated in many studies. We therefore also recommend supplementing any empirical checks with an intentional theory for why a particular intervention may or may not have a strong influence on attrition. For example, it is arguably less plausible that an intervention focused on increasing physical activity during recess would have a strong impact on attrition; on the other hand, it might be more plausible that whether a student is admitted to a charter school has a noticeable effect on whether the student chooses to enroll in a private school and hence has missing outcome data.

Strategies to Address Small Biases Due to Functional Form Misspecification in RDDs

In the case of functional form misspecification bias in RDDs, we can avoid mistakes in inference if we use existing methods to adjust impact estimates and standard errors (Calonico et al., 2014). However, this correction does increase sample size requirements. In some cases, the standard error corrections can make it practically impossible to detect impacts smaller than 0.05 standard deviations (see average minimum detectable effects in Table 5). As mentioned previously, novel approaches to RDD estimation have recently been developed that enable accurate inference while preserving statistical power (e.g., Armstrong & Kolesár, 2020; Branson et al., 2019; Calonico et al., 2020; Imbens & Wager, 2019; Noack & Rothe, 2020; Sales & Hansen, 2019). We leave to future work to investigate how these state-of-the-art methods compare to each other in similar, empirically-based education simulations.

In the meantime, we suggest that researchers always first consider whether an RDD is the most appropriate method for a particular education intervention, especially when designing *prospective* studies.[16] For example, if the relationship between the assignment variable and outcomes is likely to be highly nonlinear or if the assignment variable is very lumpy (see Appendix Figures A1 and A2 for examples), then it may be more difficult to accurately model this relationship. Appropriately hedging against this greater risk of functional form misspecification bias (e.g., by using Calonico et al.'s robust estimation) could make it difficult to detect meaningful small impacts using an RDD, even if large sample sizes are available. In these cases, to the extent they are able to, researchers might consider alternate methods for evaluating the intervention, such as other quasi-experimental designs or RCTs.

Strategies to Address Small Biases in All Study Designs

We conclude by offering a few general suggestions for researchers to consider as they plan and implement future impact studies. Our first suggestion is to reemphasize a

[16]We recognize that in the case of *retrospective* studies, researchers have much less control over the data they have and the analyses those data can support.

point made by other researchers: during the planning stages of a study, researchers should be thoughtful about what is a reasonable target minimum detectable effect for the particular intervention tested. At a minimum, researchers should consider how much the intervention might cost (other factors that might be relevant are the impacts similar interventions have obtained in the past, how impacts would compare to existing policy-relevant performance gaps, and how impacts would compare to typical academic growth trajectories). For instance, smaller impacts could still be substantively important if the cost of the intervention for the average student is relatively small; in that case, a larger sample might be appropriate in order to achieve a small minimum detectable effect. By contrast, an intervention that requires a large investment for each student served may not require a small minimum detectable effect, because the cost of implementing the intervention would only be justified if it was found to have a very large effect. Given the costs of conducting studies that are powered to detect small impacts and the increased risk of false inferences due to small biases, researchers should be able to articulate an intentional argument as to why a small impact is important to detect in each particular context.

A second, related suggestion is that regardless of the sample sizes selected, researchers should have a compelling theory of action relating proximal outcomes to distal outcomes, and they should ideally collect data on both of these outcomes. This is because a small impact on a distal outcome may be more credible if it is accompanied by a large impact on a logically connected proximal outcome. Typically, impacts are larger on proximal outcomes, and in some cases, proximal outcomes might be of intrinsic interest. For example, a text-messaging program to students might have a proximal goal of increasing attendance and a distal goal of increasing college enrollment. Because attendance itself is a behavioral outcome of interest to many schools, focusing on this outcome could allow for a more modestly powered study without sacrificing policy relevance. That said, we recognize in many cases, there is policy interest in distal outcomes such as student achievement and high school graduation. In these cases, in which studies need high statistical power to detect small distal impacts, we suggest that researchers still collect information on proximal outcomes to accompany the distal outcomes. We encourage researchers to show a strong theoretical and empirical link between the proximal and distal outcomes to help protect against potentially spurious impacts. At a minimum, if researchers find a statistically significant small impact on the distal outcome, they should be able to show that there are also larger impacts on the proximal outcome and that the proximal and distal outcomes are strongly correlated.

Ultimately, we cannot offer any single solution for addressing these challenges—the best approach is likely to vary by context. However, we do recommend that researchers resist the temptation to ignore these "small" biases. Even if these biases cannot be fully addressed, they can at least be acknowledged and mitigated to the extent possible. Consumers of research can then make more informed decisions about how much weight to put on the impact findings when making high-stakes decisions.

Acknowledgments

We thank Jessie Mazeika for excellent research assistance. We thank Luke Miratrix and two anonymous referees for helpful comments.

Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://ies.ed.gov/ncee/wwc/StudyFindings, https://ies.ed.gov/ncee/projects/evaluation/data_files.asp, https://osf.io/3x2zu/?view_only=92275575ad2c4ce5ab5cac4bdc38787a and https://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCEE20184002.

Funding

This article is based on the following report released by the Institute of Education Sciences at the U.S. Department of Education: Deke, J., Wei, T., & Kautz, T. (2017). *Asymdystopia: The threat of small biases in evaluations of education interventions that need to be powered to detect small impacts* [NCEE 2018-4002]. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. This article has been funded in part by federal funds from the U.S. Department of Education under contract [number ED-IES-12-C-0083]. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This article has been funded in part by Mathematica, Inc. (©2019 Mathematica, Inc.).

ORCID

Tim Kautz http://orcid.org/0000-0002-5631-7950

References

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness*, *3*(3), 199–253. https://doi.org/10.1080/19345741003770693

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Armstrong, T. B., & Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, *86*(2), 655–683. https://doi.org/10.3982/ECTA14434

Armstrong, T. B., & Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, *11*(1), 1–39. https://doi.org/10.3982/QE1199

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary school reading*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Barreca, A. I., Lindo, J. M., & Waddell, G. R. (2016). Heaping-induced bias in regression discontinuity designs. *Economic Inquiry*, *54*(1), 268–293. https://doi.org/10.1111/ecin.12225

Bartalotti, O. (2019). Regression discontinuity and heteroskedasticity robust standard errors: Evidence from a fixed-bandwidth approximation. *Journal of Econometric Methods*, *8*(1), 1–26. https://doi.org/10.1515/jem-2016-0007

Bartalotti, O., Calhoun, G., & He, Y. (2017). Bootstrap confidence intervals for sharp regression discontinuity designs. In M. D. Cattaneo & J. C. Escanciano (Eds.), *Advances in econometrics: Regression discontinuity designs theory and applications* (Vol. 38, pp. 421–453). Emerald Publishing Limited.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*(5), 547–556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59. https://doi.org/10.3102/0162373707299550

Branson, Z., Rischard, M., Bornn, L., & Miratrix, L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, *202*, 14–30. https://doi.org/10.1016/j.jspi.2019.01.003

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, *23*(2), 192–210. https://doi.org/10.1093/ectj/utz022

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference*, *3*(1), 1–24. https://doi.org/10.1515/jci-2013-0010

Cattaneo, M. D., Titiunik, R., & Vasquez-Bare, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of Head Start on child mortality. *Journal of Policy Analysis and Management*, *36*(3), 643–681. https://doi.org/10.1002/pam.21985

Chen, Y., Feng, S., Heckman, J. J., & Kautz, T. (2020). Sensitivity of self-reported noncognitive skills to survey administration conditions. *Proceedings of the National Academy of Sciences*, *117*(2), 931–935. https://doi.org/10.1073/pnas.1910731117

Chiang, H., Wellington, A., Hallgren, K., Speroni, C., Herrmann, M., Glazerman, S., & Constantine, J. (2015). *Evaluation of the Teacher Incentive Fund: Implementation and impacts of pay-for-performance after two years*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification, final report*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Deke, J., & Dragoset, L. (2012). *Statistical power for regression discontinuity designs in education: Empirical estimates of design effects relative to randomized controlled trials*. Mathematica Policy Research.

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

Greenberg, D., & Barnow, B. S. (2014). Flaws in evaluations of social programs: Illustrations from randomized controlled trials. *Evaluation Review*, *38*(5), 359–387. https://doi.org/10.1177/0193841X14545782

He, Y., & Bartalotti, O. (2020). Wild bootstrap for fuzzy regression discontinuity designs: Obtaining robust bias-corrected confidence intervals. *The Econometrics Journal*, *23*(2), 211–231. https://doi.org/10.1093/ectj/utaa002

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959. https://doi.org/10.1093/restud/rdr043

Imbens, G., & Wager, S. (2019). Optimized regression discontinuity designs. *The Review of Economics and Statistics*, *101*(2), 264–278. https://doi.org/10.1162/rest_a_00793

James-Burdumy, S., Deke, J., Gersten, R., Lugo-Gil, J., Newman-Gonchar, R., Dimino, J., Haymond, K., & Liu, A. Y.-H. (2012). Effectiveness of four supplemental reading comprehension interventions. *Journal of Research on Educational Effectiveness*, *5*(4), 345–383. https://doi.org/10.1080/19345747.2012.698374

James-Burdumy, S., Deke, J., Lugo-Gil, J., Carey, N., Hershey, A., Gersten, R., & Faddis, B. (2010). *Effectiveness of selected supplemental reading comprehension interventions: Findings from two student cohorts*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

James-Burdumy, S., Dynarski, M., & Deke, J. (2008). After-school program effects on behavior: Results from the 21st Century Community Learning Centers program national evaluation. *Economic Inquiry*, *46*(1), 13–18. https://doi.org/10.1111/j.1465-7295.2007.00074.x

Kane, T. J. (2015). *Frustrated with the pace of progress in education? Invest in better evidence*. The Brookings Institution.

Kolesár, M., & Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, *108*(8), 2277–2304. https://doi.org/10.1257/aer.20160945

Leamer, E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, *24*(2), 31–46. https://doi.org/10.1257/jep.24.2.31

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.

Noack, C., & Rothe, C. (2020). *Bias-aware inference in fuzzy regression discontinuity designs*. Working paper.

Pei, Z., Lee, D. S., Card, D., & Weber, A. (2020). *Local polynomial order in regression discontinuity designs*. NBER working paper 27424.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sales, A. C., & Hansen, B. B. (2019). Limitless regression discontinuity. *Journal of Educational and Behavioral Statistics*, *20*(10), 1–32.

Schochet, P. Z. (2008a). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87. https://doi.org/10.3102/1076998607302714

Schochet, P. Z. (2008b). *Technical methods report: Statistical power for regression discontinuity designs in education evaluations*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

What Works Clearinghouse. (2008). *WWC procedures and standards handbook (version 2.0)*. Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2013). *Assessing attrition bias (version 2.1)*. Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2014). *Assessing attrition bias—addendum (version 3.0)*. Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2020). *WWC standards handbook (version 4.1)*. Institute of Education Sciences, U.S. Department of Education.

Appendix

This appendix provides additional details regarding the data generating processes (DGPs) used in the RDD Monte Carlo simulations.

In our Monte Carlo simulations, we generate data using seven different DGPs. To make the simulations findings more relevant to education researchers, the DGPs are based on data from previous education studies that included math and reading post-tests and pretests. The data sources are described in Table A1.

Each DGP consists of a fifth-order polynomial equation that describes the relationship between the assignment variable (pretest) and the outcome (post-test). The coefficients in the models were estimated using the data sources described in Table A1. We report the coefficient estimates in Table A2.

Each DGP also describes the distribution of the assignment variable, including whether and how individuals are clustered within unique values of the assignment variable. These distributions were empirically estimated using the data sources described in Table A1. We report the empirical distributions in Figures A1 and A2.

Visualizations of these data generating processes are shown in Figures A3 and A4. In each figure, randomly generated data points are plotted along with the polynomials described in Table A2. The frequencies of the data points follow the empirical distributions reported in Figures A1 and A2.

Table A1. Data from past evaluations used in simulations.

Study	Purpose	Student grade	Student outcome measures	Unit of random assignment	Number of states	Number of districts	Number of schools	Number of students
Evaluation of Reading Comprehension Interventions (James-Burdumy et al., 2010)	This study evaluated the impact of four interventions on fifth-grade reading achievement.	5	Group Reading Assessment and Diagnostic Evaluation (GRADE)	School	8	10	90	6,350
Evaluation of Teacher Preparation Models (Constantine et al., 2009)	This study examined the impact of different approaches to teacher preparation on teacher practice and student performance.	K–5	Reading Comprehension, Vocabulary, and Math Concepts and Applications subtests of the California Achievement Tests, 5th Edition	Student	7	20	60	2,490
Evaluation of the Effectiveness of Reading and Mathematics Software Products (EERMSP) (Campuzano et al., 2009)	This study randomly assigned teachers to a treatment group that used a specified educational technology, or a control group that used conventional teaching approaches. The study consisted of four sub-studies of different interventions at different grade levels (see three rows below).	–	–	–	–	–	–	–
EERMSP Grade 1	–	1	Stanford Achievement Test (version 10) Reading, and Test of Word Reading Efficiency	Teacher	12	20	50	4,420
EERMSP Grade 4	–	4	Stanford Achievement Test (version 10), Reading	Teacher	9	10	40	3,110
EERMSP Grade 6	–	6	Stanford Achievement Test (version 10), Math	Teacher	7	10	30	4,260

Note: Randomized controlled trials previously completed by Mathematica for IES. Student, district, and school sample sizes are rounded to the nearest 10 in accordance with National Center for Education Statistics publication policy. State sample sizes are taken from the citations listed in the first column.

Table A2. Polynomial regression results.

Study	Test score	Number of unique pretest values	Regression coefficients						
			$b0$	$b1$	$b2$	$b3$	$b4$	$b5$	Adj-R^2
Evaluation of the effectiveness of reading and mathematics software products (Campuzano et al., 2009)	SAT-10 Reading (grade 1)	95	27.7	2.16	−0.111	0.00229	$-2.02 * 10^{-5}$	$6.57 * 10^{-8}$	0.21
	SAT-10 Reading (grade 4)	58	47.3	−3.65	0.202	−0.00516	$6.66 * 10^{-5}$	$-3.31 * 10^{-7}$	0.27
	SAT-10 Math (grade 6)	46	56.1	−6.20	0.400	−0.133	$2.37 * 10^{-4}$	$-1.71 * 10^{-6}$	0.16
Evaluation of teacher preparation models (Constantine et al., 2009)	CAT-5 Vocabulary	96	25.3	0.562	−0.031	0.00109	$-1.31 * 10^{-5}$	$5.33 * 10^{-8}$	0.18
	CAT-5 Math	99	27.5	−0.124	0.0177	−0.000177	$9.90 * 10^{-7}$	$-3.22 * 10^{-9}$	0.21
	CAT-5 Reading Comprehension	91	21.5	1.87	−0.110	0.00291	$-3.10 * 10^{-5}$	$1.15 * 10^{-7}$	0.17
Evaluation of reading comprehension interventions (James-Burdumy et al., 2010)	GRADE	31	2510	−137	2.96	−0.0311	$1.59 * 10^{-4}$	$-3.17 * 10^{-7}$	0.19

Note: Data are from the restricted-use files corresponding to the listed studies. This table reports coefficients from a regression of scores on the specified test at follow-up on a fifth-order polynomial of scores from the same test administered at baseline. Specifically, the regression is $y = b0 + b1 * x + b2 * x^2 + b3 * x^3 + b4 * x^4 + b5 * x^5$, where y is the follow-up test score and x is the baseline test score.

CAT-5: California Achievement Tests, 5th Edition; GRADE: Group Reading Assessment and Diagnostic Evaluation; SAT-10: Stanford Achievement Test (version 10).

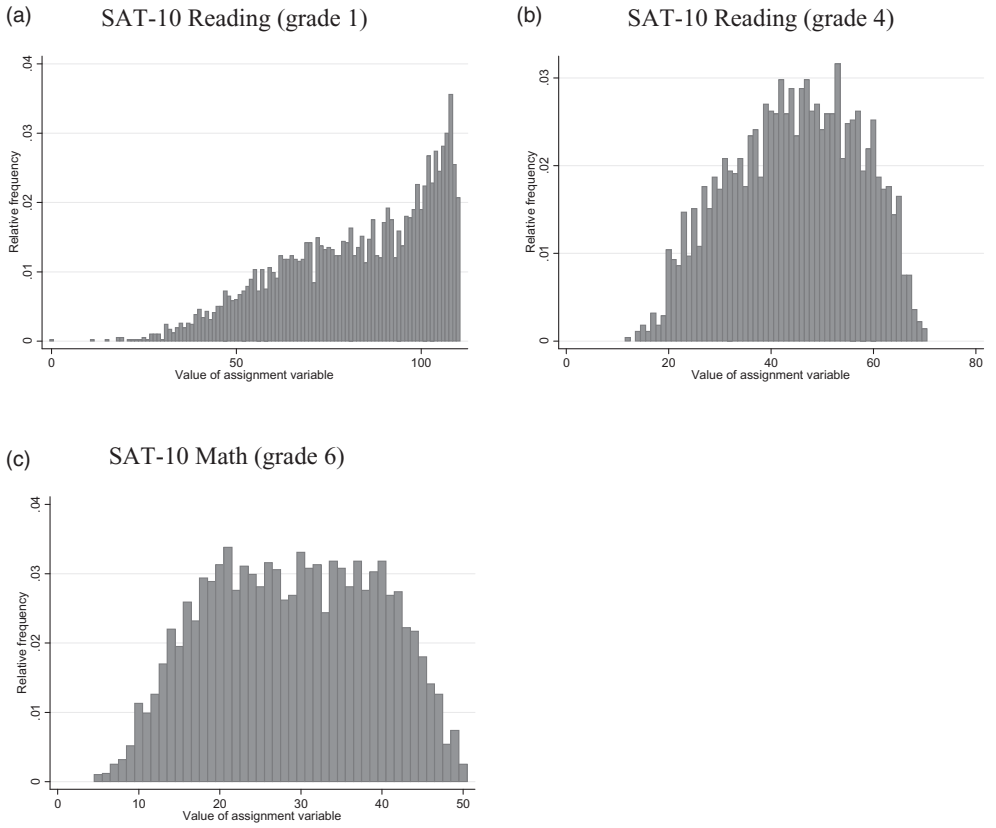

Figure A1. Relative frequency by unique value of the assignment variable (SAT-10). *Note*: Data are from the restricted-use files for Evaluation of the Effectiveness of Reading and Mathematics Software Products (Campuzano et al., 2009). This figure reports the relative frequency for each unique value of the assignment variable. SAT-10: Stanford Achievement Test (version 10)

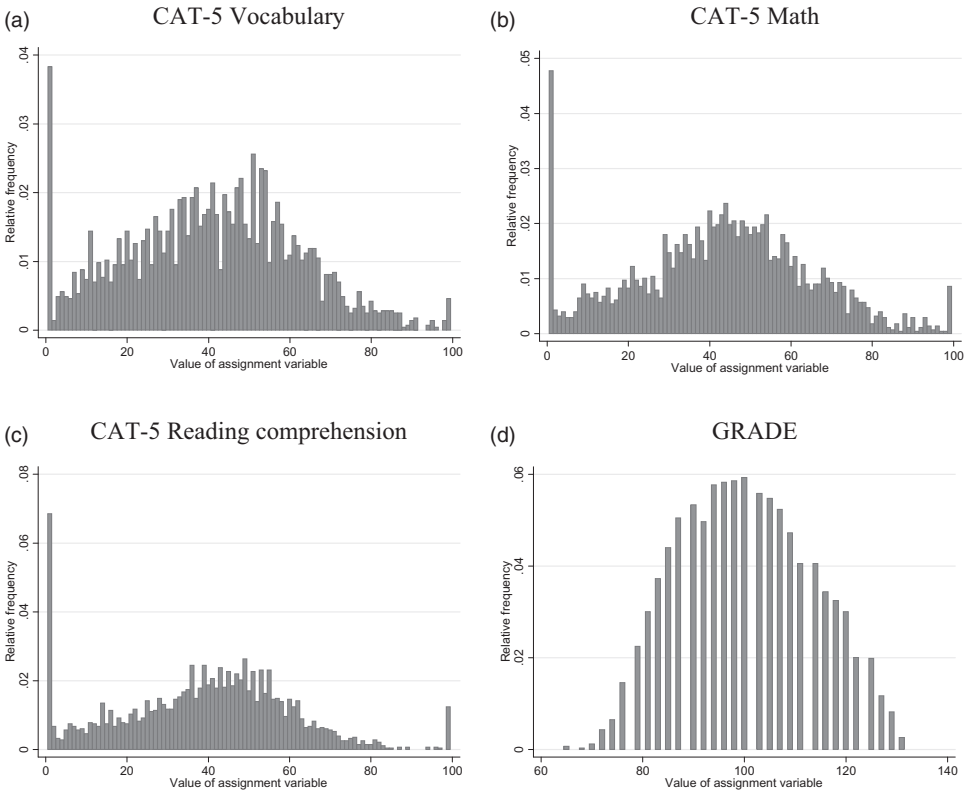

Figure A2. Relative frequency by unique value of the assignment variable (CAT-5 and GRADE). *Note*: Data are from the restricted-use files for Evaluation of Teacher Preparation Models (Constantine et al., 2009) and Evaluation of Reading Comprehension Interventions (James-Burdumy et al., 2010). This figure reports the relative frequency for each unique value of the assignment variable. CAT-5: California Achievement Tests, 5th Edition; GRADE: Group Reading Assessment and Diagnostic Evaluation.

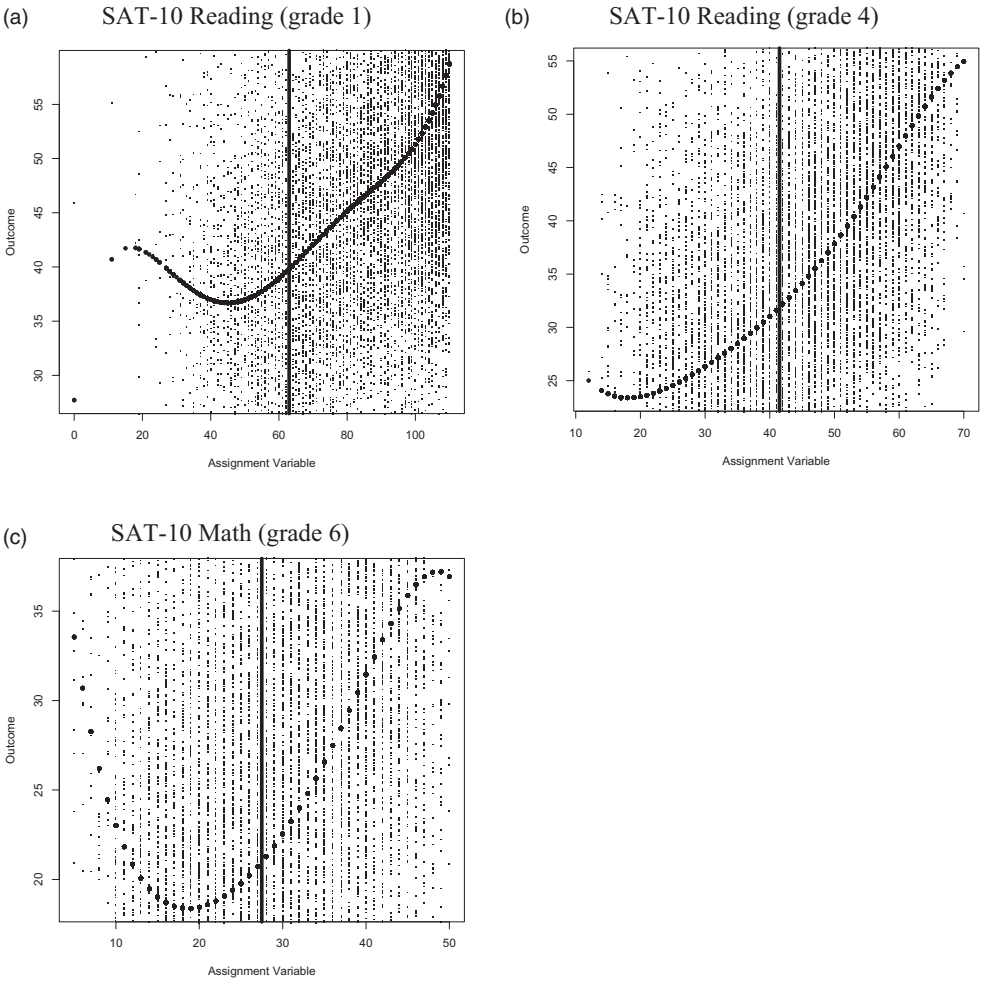

Figure A3. Visualization of data generating models for SAT-10. *Note*: Data are from the restricted-use files for Evaluation of the Effectiveness of Reading and Mathematics Software Products (Campuzano et al., 2009). This figure displays the data generating process for each outcome. SAT-10: Stanford Achievement Test (version 10).

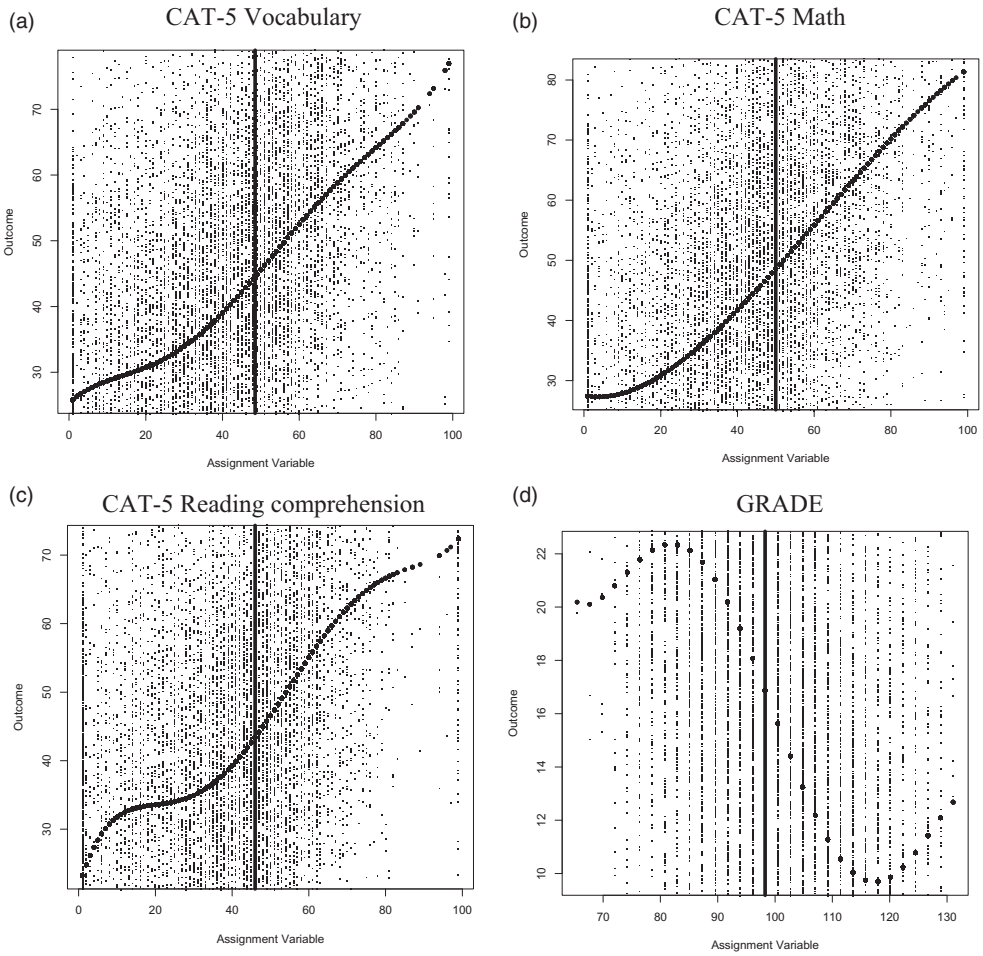

Figure A4. Visualization of data generating models for CAT-5 and GRADE. *Note*: Data are from the restricted-use files for Evaluation of Teacher Preparation Models (Constantine et al., 2009) and Evaluation of Reading Comprehension Interventions (James-Burdumy et al., 2010). This figure displays the data generating process for each outcome. CAT-5: California Achievement Tests, 5th Edition; GRADE: Group Reading Assessment and Diagnostic Evaluation.