Methods for Policy Analysis

Burt S. Barnow, Editor

THE INTERNAL AND EXTERNAL VALIDITY OF THE REGRESSION DISCONTINUITY DESIGN: A META-ANALYSIS OF 15 WITHIN-STUDY COMPARISONS

Duncan D. Chaplin, Thomas D. Cook, Jelena Zurovac, Jared S. Coopersmith, Mariel M. Finucane, Lauren N. Vollmer, and Rebecca E. Morris

Abstract

Theory predicts that regression discontinuity (RD) provides valid causal inference at the cutoff score that determines treatment assignment. One purpose of this paper is to test RD's internal validity across 15 studies. Each of them assesses the correspondence between causal estimates from an RD study and a randomized control trial (RCT) when the estimates are made at the same cutoff point where they should not differ asymptotically. However, statistical error, imperfect design implementation, and a plethora of different possible analysis options, mean that they might nonetheless differ. We test whether they do, assuming that the bias potential is greater with RDs than RCTs. A second purpose of this paper is to investigate the external validity of RD by exploring how the size of the bias estimates varies across the 15 studies, for they differ in their settings, interventions, analyses, and implementation details. Both Bayesian and frequentist meta-analysis methods show that the RD bias is below 0.01 standard deviations on average, indicating RD's high internal validity. When the study-specific estimates are shrunken to capitalize on the information the other studies provide, all the RD causal estimates fall within 0.07 standard deviations of their RCT counterparts, now indicating high external validity. With unshrunken estimates, the mean RD bias is still essentially zero, but the distribution of RD bias estimates is less tight, especially with smaller samples and when parametric RD analyses are used. © *2018 by the Association for Public Policy Analysis and Management.*

INTRODUCTION

Internal validity refers to how confident we can be that the observed association between an intervention and outcome is causal (Campbell, 1957). The epistemological warrant for valid causal inference depends on how well all alternative interpretations of the observed relationship are ruled out (Popper, 1934, 1963). To this end, the laboratory-based natural sciences use designs that capitalize on the researcher's

Journal of Policy Analysis and Management, Vol. 37, No. 2, 403–429 (2018)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22051

ability to introduce, remove, and reintroduce an intervention at will, usually at different times with different units (Cook & Campbell, 1979). Such control is less available in the policy sciences where random assignment is now preferred for ruling out alternative interpretations to the claim that two variables are *causally* related (e.g., Fisher, 1925).

External validity refers to the generalization of internally valid causal claims, and so describes the domains to which they apply (Campbell, 1957). In the applied social sciences, policymaking bodies often seek to generalize to populations of persons, settings, times, and ways of operationalizing interventions and outcomes that they specifically name and for which they are professionally accountable. In basic science, external validity has more to do with identifying causal relationships that are either universal or contingent on theoretically relevant and specifically named attributes of persons, settings, times, and both cause and effect constructs. All science aspires to high internal *and* external validity, and the present study is no exception.

One part of our study seeks to learn whether regression discontinuity (abbreviated to RD and invented by Thistlethwaite & Campbell, 1960) identifies internally valid causal relationships. We test this, not theoretically, but empirically based on the size of the difference between a local RD estimate at the cutoff and a randomized control trial (RCT) estimate made at the same point, the latter serving as the causal benchmark. No difference between the RCT and RD estimates implies perfect internal validity for the RD, while the magnitudes of the differences estimate how much bias is obtained.

This method of estimating RD bias is called the design experiment or the within-study comparison method (WSC). Its first use was in LaLonde (1986), and the first analysis of the strengths and limitations of WSC as a method is in Cook, Shadish, and Wong (2008). More than 70 WSCs have been conducted to date, including at least three in this journal (Bifulco, 2012; Wilde & Hollister, 2007; Wing & Cook, 2013). Two other studies in this journal used somewhat comparable methods (Jung & Pirog, 2014, 2017). Most WSCs test how effective design and analysis methods are for reducing the population differences that cloud the causal interpretation of nonexperimental data. In RCTs, the treatment and control groups are from the same population, whereas in nonexperiments they come from different populations and so confound treatment and population effects. Various design and analysis options have been suggested for ruling out the population difference, and WSCs test the ability of these options to reduce or eliminate the bias arising from the initial population difference. No difference between the final RCT and nonexperimental estimates is interpreted as an absence of bias, while a difference suggests bias in the nonexperiment despite the means taken to eliminate it that are therefore rated as inadequate. But such an interpretation requires holding constant all those features of the RCT and nonexperiment that might otherwise affect the study outcome so that the final bias estimate is not due to: (1) different treatments in the RCT and nonexperiment, (2) different ways of measuring the outcome in each design, (3) different quantities being estimated in each design (e.g., a local average treatment effect [LATE] in RD but an average treatment effect [ATE] in an RCT), or (4) estimation error.

This paper uses WSC methodology to estimate RD's internal validity. In asymptotic theory, both RD and RCT are expected to produce the same unbiased causal result at the treatment cutoff, as was proven first by Goldberger (1972) and then independently invented in other disciplines (for a history, see Cook, 2008). The lack of bias is also an obvious feature of all correctly implemented random assignment procedures, including at any point along a third variable that serves as an assignment variable. So, demonstrating the empirical correspondence of RD and RCT estimates at the same local cutoff point has no theoretical payoff. However, it can have a practical yield, for sampling and measurement error, and implementation and analysis shortfalls with each design, mean that perfect correspondence will not

arise in actual research practice. Important in this regard is that more shortfalls are expected with RD than RCT because of the former's lesser statistical power (Goldberger, 1972; Schochet, 2009), its need to specify functional forms or bandwidths correctly (Imbens & Lemieux, 2008), and its shorter history for identifying ways to improve design implementation. Thus, there are reasons to expect more bias with RD than RCT estimates.

Nonetheless, RCTs are still fallible WSC benchmarks, and in this context Rubin (2008) prefers RCTs with large samples and many covariates so as to reduce the odds that the final difference between RCT and nonexperimental results is due to sampling error in the RCT rather than to bias in the nonexperiment—here an RD. This dilemma is less acute when analyzing multiple WSCs, for the average RCT estimate across 15 studies will be more stable than any single RCT typically affords. We henceforth assume that the average difference between RCT and RD estimates reflects more the limitations of RD than RCT designs, thus justifying our use of the label "RD bias" to designate the final difference between causal estimates.

One purpose of this paper is to test RD's internal validity by calculating how much the average RD estimate at the cutoff differs from the average RCT estimate at this same point, doing so across 15 heterogeneous studies that each used WSC methodology and that collectively provide a more precise bias estimate than a single WSC dataset could achieve.

It is important to understand exactly what such a test of RD bias means. In RD, we use parametric or nonparametric regression methods to separately *predict* treatment and control group values of the outcome at the cutoff—data from one side of the cutoff are used for the treatment group prediction and data from the other side for the comparison group prediction. The RD effect is the difference between these predicted values and bias results because the functional forms generating the predicted values are poorly specified. Using an RCT benchmark requires estimating its impacts at the same cutoff value used in the RD. In the RCT, the treatment and control groups overlap along all of the assignment variables, and it is these overlapping data that are used to compute the benchmark estimate. This benchmark is calculated either as (1) the parametric difference between the treatment and control group predictions at the cutoff, or as (2) the nonparametric difference between treatment and control means in a small area immediately adjacent to the cutoff. The second RCT method is nonparametric as it does not require strong assumptions regarding the relationship between the running variable and outcome. Both of these RCT estimation procedures have been used to compare RCT and RD estimates, entailing RCT estimates that are unusually local so as to preclude confounding designs (RCT vs. RD) with estimands (the usual RCT ATE vs. the usual RD LATE). The RCT and RD estimation methods are not identical, the RD being more dependent on unsupported or only partially supported functional form or bandwidth assumptions when compared to either of the RCT estimation methods above. In particular the RCT treatment and control data overlap along all of the assignment variables. In contrast, in RD the treatment and control group data do not overlap at all. Thus, our meta-analytic test of RD bias can be interpreted as a test of the difference between two ways of estimating the same causal quantity at the same point on the same assignment variable, but with the RD method being theoretically more problematic because of its greater dependence on functional form or bandwidth assumptions.

How estimates of RD bias vary across the 15 WSCs will have important implications for how practical RD is as a causal design. The more tightly study-specific RD bias estimates cluster around an average close to zero, the greater is the assurance that future RD studies are likely to be unbiased and so their use can be promoted. But when the RD bias estimates are diffuse, less reliance can be placed on individual RD studies to produce unbiased results. This is obviously true when the average RD bias is large, but it even holds if the average bias is essentially zero.

So, our second purpose is to address RD's external validity by describing how RD/RCT differences are distributed across these 15 WSCs. At issue is how tight the distribution is. If they are not tightly distributed, this means trying to predict how the variation in effect sizes is related to how the 15 WSCs vary in settings, interventions, outcomes, designs, and analysis methods.

STUDY PURPOSES IN DETAIL

Among nonexperimental causal methods, RD is theoretically special because the process of selection into treatment is well understood, making it easy to model this process and thus to generate a statistically consistent causal estimate. In sharp RD, treatment allocation is completely known and perfectly measured—whether units experience the treatment or control status depends on their obtained score on the assignment variable and nothing else. So, when this score and the binary treatment indicator are included in the outcome model, their effect is to make the treatment conditionally independent of the outcome at the cutoff. In well-implemented RCTs, the treatment assignment process is also fully known, since it depends only on chance and since mechanisms for assigning by chance are not in dispute. The major contrast is with nonexperimental methods other than RD, for then the treatment assignment process is usually partially known but it is rarely fully known. It is from this only partial knowledge that the internal validity threat of omitted variable bias arises. And even if all the needed covariates were known, bias could still arise if any covariate was measured with error (Lord, 1960; Steiner, Cook, & Shadish, 2011). In contrast, the treatment assignment process in RD depends on the observed assignment covariate rather than on some unobserved latent variable, making it irrelevant to RD bias how reliably the assignment variable is measured in either sharp or fuzzy RD (Imbens & Lemieux, 2008).

The uniqueness of RD among nonexperimental methods seems counterintuitive at first because, in sharp RD, the treated and untreated units fail to overlap at every point on the assignment variable. There are no data to support the crucial assumption about missing potential outcome slopes that all RD requires—knowledge of how the assignment variable and outcome would have been related for treated units if they had not been treated. Absent such knowledge, it is customary to limit causal inference to the treatment cutoff where the treatment and comparison groups are closest to an overlap (Imbens & Lemieux, 2008). This limitation is serious, for it excludes all the other treated units for which unbiased estimates can be expected only if the treatment effect is constant across the entire treated section of the assignment variable—a heroic assumption few analysts are willing to make. RCTs have a clear advantage for causal generalization; they usually estimate the ATE for the study population rather than the LATE at the cutoff. However, to avoid confounding the RCT and RD designs with ATE and LATE estimands, three of the 15 RCTs we present estimated the RCT and RD effects at exactly the same cutoff point as the RD (Green et al., 2009; Shadish et al., 2011; Tang, Cook, & Kisbu-Sakarya, forthcoming), while the other 12 estimated their RCT LATEs as ATEs *within a narrow segment of the assignment variable immediately adjacent to the cutoff*.

The expected bias is zero in RCTs, whether for the whole population or just for those scoring at or near the cutoff. However, in RD the expectation of zero bias asymptotically holds only at the cutoff. Assuming the same cutoff in RD and RCT implies the same impact for each design and thus an expected difference of zero between them. If the difference is not zero, its size is defined as the magnitude of the estimated RD bias. One purpose of the research synthesis we present here is to estimate how closely the average difference between RD and RCT estimates approximates the predicted difference of zero.

Since RD's internal validity is not in dispute in theory, the focus in this paper is on its internal validity in research practice. RD is known for its limitations of precision (Goldberger, 1972; Schochet, 2009), implementation (Lee, 2008), and especially, analysis (Calonico, Cattaneo, & Titiunik, 2014). It requires more assumptions than the RCT, and the behavior of these assumptions is generally less well known and their diagnostic checks are less sensitive, making RD more susceptible to influence from researcher expectations that might incline study results toward what is expected instead of what is true (Ioannidis, 2005). However, RCTs are subject to sampling error, especially with smaller sample sizes and, similarly, when RCT estimates are made close to a cutoff rather than for all those treated. We follow Rubin's (2008) advice, therefore, and for each RCT code its sample size, whether covariates are used to reduce any obtained pre-intervention group differences, and whether balance tests detect implementation pitfalls. Such codes allow us to examine the extent to which obtained RCT/RD differences vary with RCT quality rather than RD bias. However, it is worth noting that error plays a smaller role in undermining the benchmark status of RCTs when syntheses of multiple RCTs are involved rather than single RCTs. The total RCT sample size is obviously larger in a synthesis and there is no obvious reason to assume that RCT errors cumulate in one causal direction versus the other. Notwithstanding, if the difference between RCT and RD estimates does not vary by RCT sample size, use of covariates, and the results of balance tests, we can be all the more confident in interpreting RCT and RD impact differences as due to RD bias more than RCT error.

Most researchers conduct single RD studies rather than syntheses. This forces them to live with whatever samples sizes and shortfalls of implementation and analysis characterize their studies. Given this reality, there is an obvious interest in RD bias estimates from individual WSC studies as well as in the average RD bias estimates that result from synthesizing across 15 WSCs. Fortunately, meta-analyses can describe the variation in study-specific RD bias estimates. The tighter this distribution is around the expected RD bias of zero, the more confident individual research sponsors, researchers, and instructors can be that future RD studies will be valid. However, if the average estimated bias is clearly nonzero, or even if it is zero but estimates of RD bias vary between studies in substantively important ways, then the empirical support for RD as a practical causal method is weakened, whatever its warrant in theory. In this research synthesis we estimate the average RD bias and also describe the distribution of individual study RD bias estimates around whatever mean difference is obtained.

LITERATURE REVIEW

The first WSC contrasting RD and RCT impacts at the cutoff was by Aiken et al. (1998). Their treatment was a remedial English writing course for students at a large state university. Students who scored below lower bounds on the ACT or SAT were required to take this course while those who scored above different but higher bounds were required to take a regular English course. Those between the two sets of bounds participated in a random assignment lottery to determine which course they would take. Those in between the bounds were about 58 percent of those taking the ACT and 18 percent of those taking the SAT. RCT estimates were calculated within these bounds, while the RD estimates were calculated at the respective upper and lower bounds. Estimates were calculated on two different tests of written English, and the RD and RCT estimates showed no clear design differences, indicating little or no RD bias. (By today's standards, the RCT effects would be calculated as LATEs at the bounds rather than as a single ATE, but that was not the case here.) The study took place at a specific time, in a specific university, and with one intervention, one

outcome construct (writing), and modest sample sizes that varied by outcome—about 100 for the RD and 144 for the RCT. Thus, the study offered promise for RD's internal validity, but the confounding of designs and estimands was an issue as was its limited external validity.

Two syntheses of WSC results exist. The first was a meta-analysis that focused on "difference-in-difference" related designs and not on RD (Glazerman, Levy, & Meyers, 2003). The second (Cook & Wong, 2008) did deal with RD's internal validity but used qualitative rather than meta-analytic methods since only three WSCs were then available—in higher education (Aiken et al., 1998), job training (Black, Galdo, & Smith, 2007), and welfare reform (Buddelmeyer & Skoufias, 2004). That review concluded there was no consistent evidence of RD bias. At least 12 more WSC tests of RD's internal validity have accumulated since, and we present them later. It is noteworthy that so many research groups have sought to test RD's internal validity in research practice despite its impeccable justification in theory. It implies there must be concern about its internal validity when implemented in field settings.

Past research is also available on how to improve WSC practice (e.g., Cook, Shadish, & Wong, 2008). WSCs vary the process of how the same treatment is assigned in order to examine how various design and analysis specifics then succeed in reducing whatever initial bias arises due to the nonexperimental source of variation in one design. This requires that the other design be unbiased and that all other factors are held constant between designs that might affect the study outcome—thus the treatment itself plus outcomes, settings, populations, time of study, and estimands. However, WSCs on RD are unique relative to WSCs on other kinds of nonexperiments. First, only an RCT can serve as a valid benchmark for an RD, whereas WSCs on other nonexperimental topics can use either an RCT or RD benchmark, given that each is unbiased asymptotically. Second, only when RD is the nonexperimental design is it known that the expected bias is zero asymptotically—at least at the cutoff. With other kinds of nonexperiment, it is impossible to know how much bias to expect.

Yet sampling error occurs in both the RCT and RD studies to be contrasted, and so we should not expect exactly zero RD bias in WSCs. This raises the question of how close to zero the obtained design difference must be in order to conclude that it falls within or beyond a region of practical equivalence and so is, or is not, worth worrying about (Wilde & Hollister, 2007). Conventions exist for asserting practical equivalence, but none is perfect—for example, the RCT and RD estimates should not statistically differ from each other, or the RCT and RD estimates should differ by less than 0.10 standard deviation units, or similar policy consequences would be drawn from each design whatever the size of the design difference. The meta-analytic context affords greater reliability and so reduces the saliency of this issue, but it does not eliminate it entirely. Whatever estimate of average RD bias is attained, we need to consider whether it is of a magnitude we should worry about.

STUDY METHODS USED TO COMPARE RD AND RCT ESTIMATES

Across the 15 WSCs, three different methods were used to create comparable RD and RCT estimates at the treatment cutoff. Shadish et al. (2011) was the only study to use a four-arm WSC method. In this, units were first randomly assigned to participate in an RCT or RD design and then, within each, they were subsequently assigned either to the treatment or control condition. How this last assignment took place depended on the first stage design to which they were allocated and so four study arms resulted from crossing the two designs with the two treatment conditions in each. All units in the RCT and RD received the same treatment or control experiences, and assessments were made in exactly the same way at both

pretest and posttest. LATE estimates were computed in both the RCT and RD and were differenced to give the estimate of RD bias.

Nine other WSCs used "a tie-breaker experiment" (Cook & Campbell, 1979). In this design, random assignment is made to treatment and control groups within one segment of the assignment variable, while the remaining observations are assigned by RD. For example, education evaluations might use baseline test scores as the assignment variable and then (1) assign all students below the 40th percentile to a remedial treatment, (2) randomly assign all students between the 40th and 60th percentiles to either the treatment or control condition, and (3) not treat any student scoring above the 60th percentile. Then, RD impacts can be computed at the 40th and 60th percentiles, and each can then be compared to RCT estimates at these same points. An alternative version of the design has just one cut point, say at the 40th percentile. Then, all students scoring below that point might be treated while all those scoring above it are randomly assigned to treatment or control status. The result is a single RD estimate at the 40th percentile that can then be compared to the RCT estimate at this same point.

The five remaining WSC studies of RD bias used a synthetic method that generates the RD data from an existing RCT dataset. The essence of the method is to designate a continuously distributed baseline variable from the RCT as the RD assignment variable—for example, baseline test scores—and to select a specific value on this as the cutoff—say, the 50th percentile. RCT control group cases from one side of this cutoff are then removed—say those below the 50th percentile, as are treatment group cases from the other side—say those above the 50th percentile. The result is an RD design with treatment data only below the cut point and control group data only above it. The LATE impact of the educational intervention can then be estimated using RD methods, and it can be compared to the same RCT estimate to determine RD bias. A noteworthy feature of such synthetic RDs is that they inevitably have fewer observations than the original RCTs because of the need to drop cases in order to construct the RDs. However, this synthetic method does enable one to use one dataset to produce many RD and corresponding RCT estimates using different running variables and cutoff points.

These three methods for constructing WSCs that estimate RD's internal validity vary in how closely they mimic what we believe to be routine RD practice. WSCs using synthetic RD construction face all the usual challenges of RD analysis but not the challenges associated with RD implementation, for no such implementation is needed. Researchers merely reconfigure existing RCT data! The four-arm WSC method is subject to both the RD implementation and analysis problems, but there is only one example of its use to date and that took place in a contrived online context with a short-acting intervention. Tie-breaker experiments are subject to all the usual challenges of RD implementation and analysis, and most uses of it to date entail two tests of RD bias, once at each corner of the assignment variable where the RCT takes place. Otherwise, the tie-breaker RD component is like any other RD, so the nine tie-breaker WSCs are the most realistic of those studies with which this meta-analysis deals. As a result, we test whether RD bias—viz., the size of the RCT/RD impact difference—varies by whether or not the WSC uses a tie-breaker design.

To compare across studies, we scale all impact estimates in standard deviation units of the original study outcomes. We do this so that design differences with a positive sign reflect RD results that are more socially desirable than RCT results. The socially desirable direction of effects was always clear—for example, toward higher academic achievement, greater earnings, or more positive health. We scale this way since it is often presumed that nonexperiments require more numerous and more opaque assumptions that leave more room for researchers' wishes, hopes, and hypotheses to affect study results (Ioannidis, 2005). So, a positive difference means

that the bias from RDs has inflated the social desirability of results compared to what an RCT would achieve.

We analyzed the pooled WSC data in two ways. Our frequentist analysis presents estimates of (1) the mean RD bias over 15 studies, (2) the confidence interval around this mean, and (3) the distribution of study-specific WSC bias estimates. A Bayesian analysis estimates similar things. The frequentist and Bayesian model results are not expected to differ much since no substantively meaningful priors are specified. Nonetheless, the Bayesian results have one advantage for point (2) above. They estimate the probability that the true bias falls within a particular range, and the practical implications of a high probability of small RD bias are quite different from the implications of a low probability of such bias or of any probability of large bias. By contrast, the interpretation of frequentist confidence intervals is somewhat more convoluted, though they are presented too.

SUMMARY OF PURPOSES AND GENERAL METHOD

This paper presents two types of meta-analysis, each of 15 WSCs that contrast RCT and RD causal estimates at the treatment cutoff. We calculate (1) the average difference between the RCT and RD estimates, interpreting the size of this difference as the estimated RD bias. We also calculate (2) how the study-specific RD bias estimates are distributed around the average bias. Asymptotically, the average estimated bias should be zero and the study-specific effect sizes should not differ. However, sampling error and imperfectly conducted RD studies justify an empirical check of whether the obtained bias in RD is *robustly* minimal in practice. That is, we ask: Is RD really just like an RCT at the cutoff, not just in the theoretical world of abstract expectations, but in the concrete world of research practice? The value of basic RD for research practice is indicated if the average RD bias is close to zero and the study-specific differences are tightly distributed around this average. Conversely, RD is less useful if the average difference is not zero or the distribution of study-specific bias estimates is highly variable.

METHOD FOR META-ANALYZING WSC ESTIMATES

This study meta-analyzes WSCs that deliberately sought to identify whether RCTs and RDs produce similar causal estimates, given that the two designs have the same treatment group, the same estimand, and are otherwise similar on as many outcome-correlated irrelevancies as possible, including data collection methods. WSCs assess how successful various design and analysis procedures are in reducing the bias that arises in nonexperiments because of the absence of a full description of how the treatment and comparison group populations differ. RCTs form their contrast groups at random, and so no bias is expected; but nonexperiments, including RDs, have contrast groups from different populations and so there is a potential for bias. The final difference between RCT and RD estimates we interpret as due to estimated RD bias. This interpretation requires that there be zero bias in the RCT, not just in expectation, but also in practice. It is fortunate, then, that any bias from RCT practice should be less after pooling across 15 RCTs as opposed to relying on a single RCT. Even so, we estimate how RD bias varies with the sample size of each RCT and the use of covariates and balance tests. The less it varies with these factors, the easier it is to assume that average RCT/RD differences are due more to bias in the RDs than the RCTs.

Search Method

Our search for relevant studies is part of a larger effort to identify all WSCs that assess how effective are methods designed to minimize or eliminate the bias from

Table 1. Within-study comparisons of RD using RCT benchmarks.

Authors/Year	Contrasts	Field	Intervention
Aiken et al. (1998)	4	Education	Remedial education in college
Ashworth and Pullen (2015)	3	Education	Remedial English, first grade
Barrera-Osorio, Filmer, and McIntyre (2014)	8	Education	Grade 4 scholarships
Berk et al. (2010)	3	Crime	Parole services
Black, Galdo, and Smith (2007)	84	Labor	Reemployment services
Buddelmeyer and Skoufias (2004)	8	Education	Cash to mother for child attendance
Gleason, Resch, and Berk (2012)	12	Education	Teach for America and technology
Green et al. (2009)	6	Politics	Get out the vote
Hyytinen et al. (2009)	4	Politics	Incumbency advantage
Kisbu-Sakarya, Cook, and Tang (in press)	2	Education	College math and vocabulary
Moss, Yeaton, and Lloyd (2014)	4	Education	Developmental math
Nickerson (2007)	1	Politics	Get out the vote
Shadish et al. (2011)	8	Education	College math and vocabulary
Tang, Cook, and Kisbu-Sakarya (forthcoming)	6	Education	Head start
Wing and Cook (2013)	18	Welfare	Cash and counseling

Note: Each contrast is an RD impact estimate minus an RCT estimate of the same parameter. The numbers of contrasts reported here are those we used in our analyses and not the totals from these studies.

comparing treatment and comparison groups from different populations. This effort includes, not just RD, but also the use of interrupted time series and simpler "difference-in-difference" designs. By March of 2015 we had identified over 60 such studies. We then contacted about 140 researchers, including all authors of the identified papers as well as other researchers doing related work. We asked them if they could let us know about any studies they or others might be doing in the WSC area, particularly unpublished work. Through this, and routine journal search efforts, we identified about 20 more WSCs for a total of over 80, of which 15 test the efficacy of RD. We rejected WSCs in clinical medicine rather than social science (e.g., Baker & Lindeman, 2001; Cooper et al., 1997; Dahabreh et al., 2012; Deeks et al., 2003), and also papers that involved the comparison of designs across unique datasets that do not control for the confounding factors to which WSCs aspire (e.g., Dahabreh et al., 2012; Heckman, LaLonde, & Smith, 1999; Weisburd, Lum, & Petrosino, 2001). Of the 15 WSCs dealing with RD, seven were published in peer-reviewed journals, and eight were not, suggesting that we had penetrated some way into the "file-drawer" (Rosenthal, 1979). One paper even existed only in PowerPoint form (Nickerson, 2007), and so we obtained the data from the author and reanalyzed them for this paper. The reanalysis details are in our online Appendix.[1]

Table 1 indicates that the 15 WSCs vary in a number of substantively important ways. Nine focus on education but cover the gamut from pre-K to college; three deal with voting behavior; while one each is devoted to topics in criminology, labor, and

[1] The appendix is available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

welfare. Five types of assignment variables are used across the studies—test scores, household income, individual age, prior election results, and predicted length of time in social services. Respondent ages range from 3 (Tang, Cook, & Kisbu-Sakarya, forthcoming) to 95 (Wing & Cook, 2013). Some studies are very local, while others are state or nationwide in scope. Publication dates start in 1998 and continue to 2016. Such heterogeneity suggests we can probe how robust the correspondence is between RCT and RD estimates and so escape the doubts about replicability that bedevil most individual studies, including single WSCs.

However, two sources of homogeneity stand out despite the diversity—the studies tend to focus on topics of interest to applied rather than basic social science, and the sample is necessarily limited to topics where both an RCT and RD are possible. The first reflects our own interest, and so we do not consider it a problem. However, the second is potentially a much bigger problem. Why learn what reduces bias in a context where you could do an RCT anyway? Later, we discuss in detail this limit to the external validity of this synthesis presented here.

Coding Process

This is the first meta-analysis to deal with RD, and we had to develop a coding manual virtually from scratch. We first created a draft manual that was revised after the current authors separately coded and jointly discussed one especially complicated WSC—Buddelmeyer and Skoufias (2004). The remaining WSC studies were then separately coded by two study authors and, where they agreed, their common score was retained. Where they disagreed, they reconciled their coding differences. In the few cases where they could not reach a reconciliation, a third author met with them to reconcile their difference. This resulted in minor modifications being made to the codebook throughout the coding process as experience with recondite RCT, RD, and WSC issues accumulated. Eventually, we had complete and comparable data for all 15 WSCs.

Study Contrasts

The unit of analysis in this meta-analysis is the treatment/control contrast, each being the difference between an impact estimate from an RD and one from an RCT. Mostly, the number of contrasts per WSC depends on how many outcomes, posttest time points, subgroups, and analysis methods are used per study. We used the following rules to limit the coding burden and the correlation between within-study outcomes, while maximizing the within-study variation in variables of special interest. First, in the few WSCs with more than two posttest time points we analyzed the first and last post-treatment time points only. Second, we always included contrasts based on the total population studied, omitting contrasts based on subgroup impacts unless they were the only ones reported. Third, we rejected all contrasts described in the text as sensitivity tests, as well as those relegated to a study appendix, thus preferring the estimates presented in the main study tables or text. Fourth, when possible we also rejected contrasts that failed to meet criteria of WSC adequacy, including that the treatment groups be the same (or can be used to estimate the same population parameter) and that the randomization process be acceptable. However, two studies (Aiken et al., 1998; Buddelmeyer & Skoufias, 2004) had RCT and RD treatment groups that overlapped considerably but were not totally identical, and here we included the contrasts the authors preferred. This process resulted in a meta-analysis with 171 contrasts over 15 studies. Table 1 shows that the number of contrasts varied considerably across studies. The maximum was 84 (Black, Galdo, & Smith, 2007), the next highest was 18 (Wing & Cook, 2013), and the lowest was one (Nickerson, 2007). More than half of the studies had between two and six contrasts.

As a result, the impact models we used include the natural log of the number of study-specific contrasts to see if it is related to the size of the RD bias.

More important, though, is that contrasts are not likely to be independent within studies and that the amount of dependence can vary by study. To achieve a study-specific estimate of RD bias requires us to average estimates within a WSC. Meta-analytic practice has been to reject the individual contrast estimates and to use the mean of all contrast estimates within a study as the sole indicator of an effect for that study. But this ignores any signal that the individual contrasts might reveal about sources of variation in bias estimates. Instead, we estimate study-specific bias using a random effects model. The random effects model allows for covariance in contrast errors. However, it is based on the assumption that the covariance terms are the same across studies. This might or might not be true and could affect estimated standard errors in the model. As an antidote, Hedges, Tipton, and Johnson (2010) have proposed a method based on robust standard errors that makes fewer distributional assumptions. But the available evidence suggests it does not work well when synthesizing fewer than 20 studies and the present study has 15.

Outcome Variables

To permit between-study comparisons, all outcomes were transformed into contrast-specific standardized bias estimates. Six studies provided us with standardized RD bias estimates directly or with the pooled standard deviation of the original study outcome necessary to calculate it directly (Aiken et al., 1998; Gleason, Resch, & Berk, 2012; Kisbu-Sakarya, Cook, & Tang, in press; Nickerson, 2007; Shadish et al., 2011; Tang, Cook, & Kisbu-Sakarya, forthcoming). Four studies used binary outcomes (Barrera-Osorio, Filmer, & McIntyre, 2014; Berk et al., 2010; Green et al., 2009; Hyytinen et al., 2009). While it is common to use log odds ratios for binary outcomes in meta-analyses, for ease of interpretation we stuck with mean differences of the binary outcomes and calculated treatment and control group standard deviations for them as the square roots of their variances. Each variance, in turn, was calculated using the group mean of the outcome times one minus the mean. For two papers with continuous outcomes we found information about the standard deviations in prior publications (Ashworth & Pullen, 2015; Barrera-Osorio, Filmer, & McIntyre, 2014). In another two WSCs we used baseline standard deviation values or values for the full RCT sample instead of the LATE subset within this sample (Black, Galdo, & Smith, 2007; Buddelmeyer & Skoufias, 2004). In the remaining two papers, we used the ratio of the standardized and unstandardized standard errors where the former was approximated from the sample sizes in the treatment and control groups (Moss, Yeaton, & Lloyd, 2014; Wing & Cook, 2013).

In computing effect sizes we used the analytic sample sizes where provided, but used the originally assigned sample sizes if the analytic sample sizes were not presented. As noted earlier, the difference between RCT and RD estimates was scaled so that a positive difference reflects RD bias operating in the direction social policy typically aspires to affect—for example, greater academic achievement, higher incomes, or better health. Thus, an average RD bias of 0.10 indicates that the RD studies inflated positive program effects by this amount relative to the corresponding RCT estimates.

Control Variables

The potential for confounding arises when a poorly implemented RD is compared to a well-implemented RCT, for this confounds design difference and variation in implementation quality. For each contrast in each study we include a number of RD quality measures based on the analysis of Imbens and Lemieux (2008), though

additional standards are plausible based on fine detail about nonparametric analyses (e.g., Calonico, Cattaneo, & Titiunik, 2014). Specifically, we coded: (1) the natural log of the RD sample size; (2) whether the control for the assignment variable was nonparametric versus parametric under the assumption that the former is generally superior—32 percent of all contrasts are coded as nonparametric; (3) whether there was a density test for manipulation of cases close to the cutoff—yes for 31 percent of the WSC contrasts; and (4) whether optimal methods were used to select the functional form/bandwidth for the assignment variable—yes for 87 percent; we also assessed (5) whether the RD estimated a LATE at the cutoff, but this was eventually omitted because 99 percent of all contrasts were coded as yes. We sought to assess how these RD quality features were associated with the level of bias reduction achieved, expecting less bias with technically superior RDs.

We also measured and analyzed a number of measures of RCT quality: (1) the natural log of the RCT sample size; (2) whether baseline differences were reported and were small—95 percent yes; (3) whether a pretest or proxy pretest was included as a covariate in the impact analysis—19 percent of contrasts coded as yes; and (4) whether other demographic control variables were included—48 percent yes. The last two speak to the use of covariates to adjust for imbalances when using finite sample RCTs for WSC purposes; we also checked (5) whether each study reported using a valid random assignment procedure, but no study reported an incorrect procedure, so we dropped this item. Unfortunately, few studies provided data to calculate attrition rates, whether overall or differentially by treatment status, and so we could not include attrition in our analysis of RCT quality.

We also collected data on WSC quality using the six quality criteria in Cook, Shadish, and Wong (2008). But we could not use data on all of them for want of variation. For example, only two contrasts involved the use of independent researchers to analyze the RD and RCT data (both in Shadish et al., 2011), and that study was also the only one to randomly assign units to the RCT or RD before assigning them to treatment condition. However, we did include an indicator for whether or not the RD and RCT used the same estimand, though variation was limited here too since the estimands were identical in 95 percent of the contrasts.

The impact analyses also included whether the RD is the product of a tie-breaker experiment rather than a synthetic design or a four-armed study. Forty-four contrasts from five studies involved synthetic RD (Gleason, Resch, & Berk, 2012; Green et al., 2009; Kisbu-Sakarya, Cook, & Tang, in press; Tang, Cook, & Kisbu-Sakarya, forthcoming; Wing & Cook, 2013); eight contrasts came from the one four-armed WSC study (Shadish et al., 2011); and the remaining 119 contrasts came from the nine studies that used a tie-breaker design to embed the RD within the RCT.

In total, 11 covariates were used across 171 contrasts from just 15 WSCs. It is important to note how the 11 control variables were related to the study and contrast levels of analysis. Four of the controls varied by study only—use of a pretest measure of the study outcome, use of a McCrary-like density test in the RD, the log of the number of contrasts, and whether a tie-breaker WSC design was used. Another three varied within only one of the 15 studies, making them effectively study-level controls—viz., use of a baseline equivalence test, use of covariates other than the pretest, and whether the same estimand was used in the RD and RCT. Another covariate varied within only two studies—viz., whether an optimal model specification was used. This pattern of variation means that only three covariates varied within five or more WSCs, making them the only plausible contrast-level controls—viz., use of a parametric versus nonparametric RD test, the log of the RD sample size, and the log of the RCT sample size. So for the 11 covariates there are more like 15 degrees of freedom rather than 171 *independent* contrasts. Indeed, some covariates are highly correlated with each other even at the study level—for example, the log of the RD sample size correlates 0.75 with whether the average contrast is

nonparametric and correlates 0.72 with the log of the RCT sample size. We use the control variables to examine how each is related to RD bias, but their standard error estimates are affected by multicollinearity and the high ratio of predictors to operative degrees of freedom.

Estimation Model

Most meta-analyses in the social sciences use frequentist statistical frameworks that place a priority on standard errors and *P*-values. Bayesian meta-analyses are increasingly common, though more so in medicine than the social sciences (Gelman, Hill, & Yajima, 2012). With as little information about substantive priors as here, it is unrealistic to expect frequentist and Bayesian analyses to produce markedly different estimates of average RD bias and the distribution of such bias. Nonetheless, we conduct both kinds of analysis because the frequentist model is best known but the Bayesian techniques provide a more intuitive summary that is based on the probability that the difference between the RCT and RD results falls within a given range as opposed to the probability that the observed coefficient, or a larger one, would be observed if the true coefficient was zero.

Using random effects to estimate bias by study (in a Bayesian or frequentist model) offers the opportunity to take advantage of information from other studies when estimating the average bias in a given study. This process, also called "borrowing strength" or "shrinkage," begins with the assumption that the study-specific biases are related to some degree. The data dictate the extent to which that assumed relationship holds. Studies with more unexplained variation in their bias estimates are shrunken more toward the overall average while those with less unexplained variation get less shrinkage. In addition, if the data determine that standardized bias is similar across studies, the model will draw more heavily on information from other studies to estimate the bias in a single study. If the data determine that standardized bias varies widely by study, the model will draw more heavily on the data each study provides. As a result, random effects models gain additional precision in a data-driven way. This additional precision comes at a cost; combining the data from each study with information from other studies biases the estimates. Shrinkage balances the trade-off between bias and precision by minimizing the mean squared error which depends on both (Efron & Morris, 1977). We also present unshrunken results by study in Table 5 and Figure 4.

Our models have a few other key features. We mean-center the covariates so that the intercept represents the average standardized bias for an average study. The random effects models (Bayesian and frequentist) can be thought of as providing approximate inverse variance weights. This is because those models effectively down weight contrasts with higher variance estimates. We used the RD sample size to approximate the variance of the estimation error for the individual contrasts because only six of the 15 studies provided standard errors of the RD bias estimates. The frequentist confidence intervals are calculated using the Wald normal approximation, whereas uncertainty intervals for the Bayesian models are empirical, based on the relevant quantiles of the posterior probability distribution. All models have the following form:

$$C_{ij} = \alpha + \beta X_{ij} + \xi_j + \varepsilon_{ij},$$

where:

C_{ij} = contrast i from study j;
α = the intercept which estimates the mean bias;
X_{ij} = omitted in models without controls, otherwise these are the covariates described above for models with controls;

Table 2. Prior distributions for Bayesian and frequentist meta-analysis models.

Parameter	Bayesian prior	Frequentist prior
α	Normal(0, 100)	Uniform($-\infty$, ∞)
β (All components)	Normal(0, 100)	Uniform($-\infty$, ∞)
ξ_j	Normal(0, σ_ξ^2)	Normal(0, σ_ξ^2)
ε_{ij}	Normal (0, $\tau_\varepsilon^2 + \frac{c}{N_{ij}}$)	Normal(0, $\tau_\varepsilon^2 + \frac{c}{N_{ij}}$)
σ_ξ	Half-Cauchy(0, 6.25)	Fixed
τ_ε	Half-Cauchy(0, 6.25)	Fixed
c	Normal(0, 100)	Fixed

ξ_j = a random study effect; and
ε_{ij} = a contrast-specific error term.

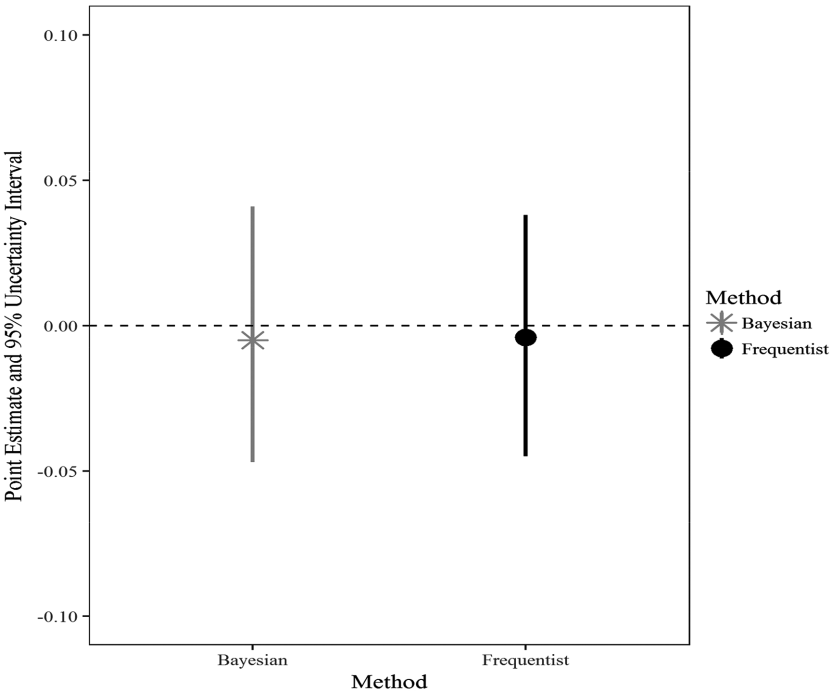
The Bayesian analysis requires prior distributions for the parameters α, β, ξ_j, and ε_{ij}. Our analysis specified the following priors:

- $\alpha \sim$ Normal(0, 100);
- each element of $\beta \sim$ Normal(0, 100);
- $\xi_j \sim$ Normal(0, ${\sigma_\xi}^2$);
- $\varepsilon_{ij} \sim$ Normal(0, $\tau_\varepsilon^2 + \frac{c}{N_{ij}}$), where τ_ε is an overall error term, N_{ij} is the RD sample size for contrast i in study j, and c is a constant estimated from the data;
- $\sigma_\xi \sim$ half-Cauchy(0, 6.25);
- $\tau_\varepsilon \sim$ half-Cauchy(0, 6.25); and
- $c \sim$ Normal(0, 100).

Three aspects of the foregoing deserve commentary. First, the variance of the error term, ε_{ij}, includes one term, τ_ε, that captures variability in true bias across contrasts within a study and another term, $\frac{c}{N_{ij}}$, that approximates the squared standard error of the standardized bias in contrast i in study j. The true standard error of the standardized bias depends on many other factors, including the RCT sample size, the use of control variables, the degree of clustering, and the degree of overlap in the RD and RCT treatment groups. We made an attempt to incorporate this additional information into our computation of this component, but the resulting weights varied too much to be plausible. So, without examining the final bias estimates, we opted for the streamlined approximation as $\frac{c}{N_{ij}}$. In results not presented here, we used the implausible weights. The RD bias estimates were not much different from those we present below.

Second, the frequentist and Bayesian estimates come from very similar models. However, the frequentist model does not account for uncertainty in the hyperparameters since traditional statistical packages cannot fit such a model while simultaneously estimating c from the data. Only the Bayesian estimation allows this. So, to retain this component in our frequentist model we used information from the corresponding Bayesian models. Specifically, we fixed the variance components (τ_ε, σ_ξ, c) at their estimates from the analogous Bayesian model and put flat priors on α and all βs. We tabulate the priors for the frequentist and Bayesian analyses in Table 2.

Third, our use of random study effects means that we control for clustering by paper, but not by dataset or study team. Clustering by dataset is not likely to matter, because no dataset is used more than once across the papers and only Gleason, Resch, and Berk (2012) used more than one dataset—they used two. No pair of papers had the same authors, and so there is no clear way to adjust for clustering



Note: Results from models without control variables reported in Table 3.

Figure 1. Estimates of Average RD Bias: Frequentist and Bayesian Results.

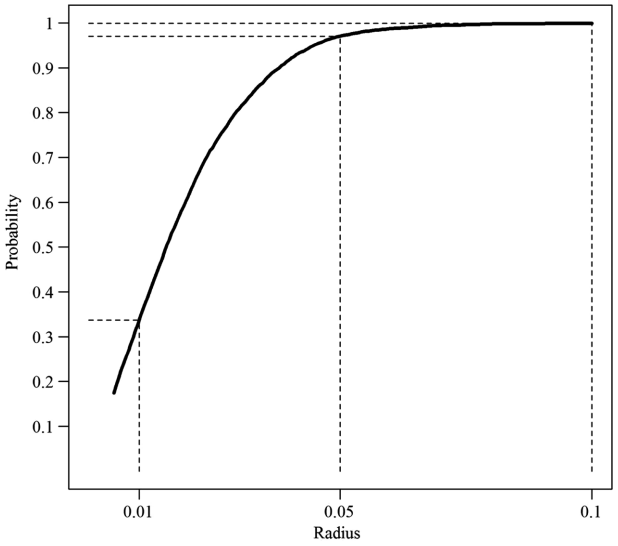
by study team. However, there is some clustering by training background, and four studies are associated with research done at Northwestern University. We do not adjust for such clustering but do examine how unique these four studies are in their RD bias estimates.

RESULTS

Average RD Bias

Figure 1 shows that the average bias is below 0.01 standard deviations in both the frequentist and Bayesian models when no between-study control variables are used. The frequentist 95 percent confidence interval ranges from −0.045 to +0.037, and the Bayesian credible interval ranges from −0.050 to +0.040. Figure 2, based on the Bayesian model, indicates a 35 percent chance that the average bias is between −0.01 and 0.01 and a 97 percent chance that it is within 0.05 of either side of zero.

Table 3 shows that adding the control variables makes no substantive change to the estimate of average RD bias. However, the standard errors with controls are three or four times higher than without them, probably due to collinearity and having 11 covariates but only 15 WSCs in a situation where many control variables do not vary, or hardly vary, by contrast within a study. Table 4 shows that none of the control variables is reliably related to RD bias in any type of analysis. The best single predictor is whether the RD contrast did or did not use an optimal functional form procedure—an explicit optimal bandwidth selection procedure in



Note: Results from Bayesian model without control variables reported in Table 3.

Figure 2. Probability of Bias around Radii from 0 to 0.10.

Table 3. Average bias by statistical model and analytic method.

Mean	Standard error	Model	Method
−0.005	0.023	Without controls	Bayesian
−0.004	0.021	Without controls	Frequentist
0.008	0.074	With controls	Bayesian
0.011	0.065	With controls	Frequentist

Note: The control variables used in lines 3 and 4 of this table are listed in Table 4. Also all models included random effects for each study.

Table 4. Bias estimates for covariates: Bayesian and frequentist models..

	Bayesian		Frequentist	
Parameter	Estimate	Standard error	Estimate	Standard error
Nonparametric control	−0.010	0.033	−0.011	0.033
McCrary test	0.052	0.093	0.050	0.085
Optimal functional form	−0.086	0.085	−0.094	0.071
Pretest	−0.030	0.109	−0.029	0.101
Demographic controls	−0.008	0.030	−0.007	0.030
Small baseline diffs	−0.015	0.089	−0.017	0.090
Same estimand	−0.022	0.107	−0.027	0.110
log (Contrasts)	0.003	0.047	0.006	0.043
log (*N* RD)	−0.019	0.016	−0.019	0.016
log (*N* RCT)	0.022	0.033	0.022	0.031
Synthetic or four-armed	−0.011	0.107	−0.014	0.099

Table 5. Study-level sample sizes and bias estimates with and without shrinkage.

	Average contrast		Average sample size	
Study	Without shrinkage	With shrinkage	RD	RCT
Aiken et al. (1998)	−0.113	−0.008	117	95
Ashworth and Pullen (2015)	0.114	0.009	164	89
Barrera-Osorio, Filmer, and McIntyre (2014)	0.001	−0.016	552	452
Berk et al. (2010)	0.027	0.010	1,800	1,559
Black, Galdo, and Smith (2007)	−0.071	−0.067	5,689	1,981
Buddelmeyer and Skoufias (2004)	−0.032	−0.018	4,963	4,072
Gleason, Resch, and Berk (2012)	−0.017	−0.005	1,343	1,745
Green et al. (2009)	−0.011	−0.004	8,950	30,038
Hyytinen et al. (2009)	−0.008	−0.002	23,203	1,351
Kisbu-Sakarya, Cook, and Tang (in press)	−0.250	−0.010	118	235
Moss, Yeaton, and Lloyd (2014)	−0.031	−0.010	2,059	59
Nickerson (2007)	−0.024	−0.006	21,522	6,250
Shadish et al. (2011)	0.156	0.050	380	189
Tang, Cook, and Kisbu-Sakarya (forthcoming)	0.027	0.014	1,098	2,449
Wing and Cook (2013)	0.095	0.064	929	1,850

Note: Estimates with shrinkage are from the Bayesian model without control variables.

nonparametric tests or a statistical method of choosing functional forms in parametric tests. But this coefficient is never significant and is less than 0.10 in magnitude, warranting the conclusion that none of these covariates significantly moderates RD bias.

The Distribution of RD Bias Estimates across WSC Studies

The first column of Table 5 reports the estimated bias for each WSC after averaging across all its contrasts and without taking account of the shrinkage that the 14 other WSCs make possible. These unshrunken results provide a picture of what researchers doing single WSCs on RD discover. The absolute values of the unshrunken RD bias estimates range from as high as 0.25 in Kisbu-Sakarya, Cook, and Tang (in press)—a small sample stress test of RD's internal validity—to a low of 0.001 in Barrera-Osorio, Filmer, and McIntyre (2014). Four of the 15 unshrunken estimates are larger than 0.10 in magnitude, implying meaningful RD bias if 0.10 standard deviations were to be accepted as the criterion for indicating no practical difference between RD and RCT estimates.

However, meta-analysis enables use of the information afforded by the other 14 WSCs too. The second column of Table 5 reports the results from a Bayesian model without controls after using study-specific random effects to "shrink" each bias estimate toward the overall mean and thus enhancing the precision of the study-specific RD bias estimates. Each shrunken estimate is a weighted average of the study-specific and of the overall average bias, with the weights chosen to minimize mean squared error. The absolute bias ranges now from 0.067 to 0.002, as opposed to from 0.250 to 0.001 for the corresponding unshrunken estimates; and it is also now evident that differences between shrunken and unshrunken estimates tends to be larger with smaller RD sample sizes. This is particularly evident with

Table 6. Shrunken study-level bias estimates (standard errors).

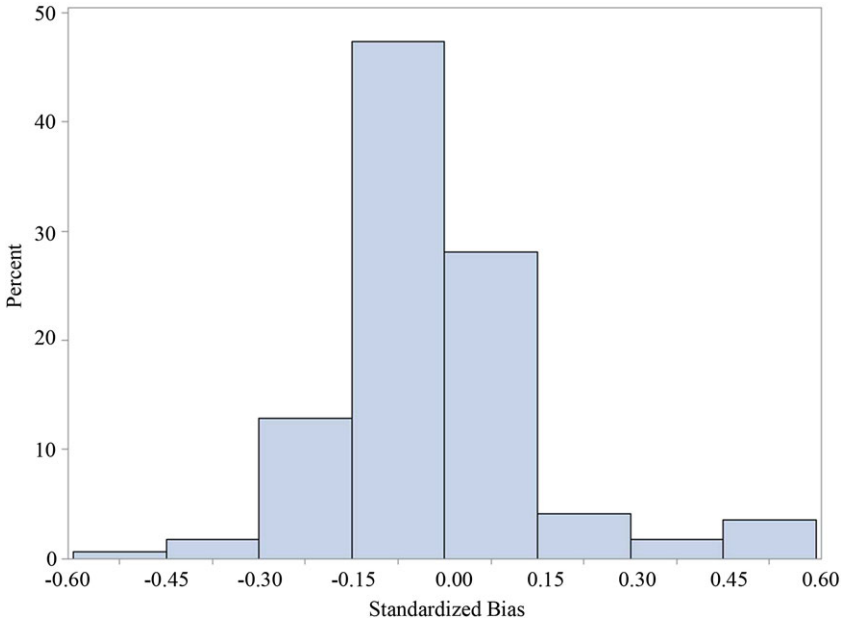
	Without controls		With controls	
Study	Bayesian	Frequentist	Bayesian	Frequentist
Aiken et al. (1998)	−0.008	−0.007	−0.076	−0.082
	(0.055)	(0.054)	(0.122)	(0.108)
Ashworth and Pullen (2015)	0.009	0.010	0.053	0.053
	(0.055)	(0.053)	(0.109)	(0.107)
Barrera-Osorio, Filmer, and McIntyre (2014)	−0.016	−0.017	−0.053	−0.056
	(0.043)	(0.044)	(0.087)	(0.081)
Berk et al. (2010)	0.010	0.011	0.037	0.040
	(0.047)	(0.045)	(0.106)	(0.104)
Black, Galdo, and Smith (2007)	**−0.067**	**−0.069**	−0.034	−0.042
	(0.025)	**(0.023)**	(0.132)	(0.116)
Buddelmeyer and Skoufias (2004)	−0.018	−0.019	−0.108	−0.116
	(0.034)	(0.033)	(0.112)	(0.094)
Gleason, Resch, and Berk (2012)	−0.005	−0.005	0.006	0.006
	(0.036)	(0.037)	(0.104)	(0.095)
Green et al. (2009)	−0.004	−0.004	0.005	0.006
	(0.035)	(0.035)	(0.104)	(0.100)
Hyytinen et al. (2009)	−0.002	−0.002	0.027	0.029
	(0.037)	(0.038)	(0.094)	(0.090)
Kisbu-Sakarya, Cook, and Tang (in press)	−0.010	−0.008	−0.065	−0.071
	(0.057)	(0.053)	(0.117)	(0.108)
Moss, Yeaton, and Lloyd (2014)	−0.010	−0.011	−0.022	−0.027
	(0.042)	(0.041)	(0.106)	(0.106)
Nickerson (2007)	−0.006	−0.004	0.035	0.039
	(0.048)	(0.049)	(0.121)	(0.114)
Shadish et al. (2011)	0.050	0.049	0.148	0.158
	(0.050)	(0.045)	(0.109)	(0.092)
Tang, Cook, and Kisbu-Sakarya (forthcoming)	0.014	0.014	−0.048	−0.050
	(0.041)	(0.042)	(0.111)	(0.098)
Wing and Cook (2013)	0.064	**0.069**	0.110	0.114
	(0.038)	**(0.034)**	(0.102)	(0.093)

Note: Bold indicates statistically significant at the 0.05 level.

Kisbu-Sakarya, Cook, and Tang (in press) where the original estimate of 0.250 shrinks to 0.001.

Table 6 reports shrunken RD bias estimates from both the Bayesian and frequentist analyses. Without the controls, where there is no reason to expect variance inflation, all shrunken bias estimates are under 0.07 and, to the eye, their distribution seems leptokurtic. In the frequentist model, two estimates are statistically significant, but each is below 0.10 in magnitude and only one of them is also significant in the Bayesian model. This is little more than would be expected by chance; and even if they were not due to chance, the estimated RD bias remains small in magnitude.

Standard errors increase when the 11 covariates are included in the impact models, given the multicollinearity. Now, none of the shrunken effects indicates a reliable difference between the RCT and RD estimates, though three of the RD bias estimates are larger than 0.10—viz., 0.148, 0.110, and 0.108. The other 12 estimates cluster between 0.08 and zero. It seems, then, that individual bias estimates do not differ much in analyses without controls, but a few do when control variables are used that inflate rather than reduce standard errors.



Note: This is the distribution of the standardized contrasts used in this study, without shrinkage.

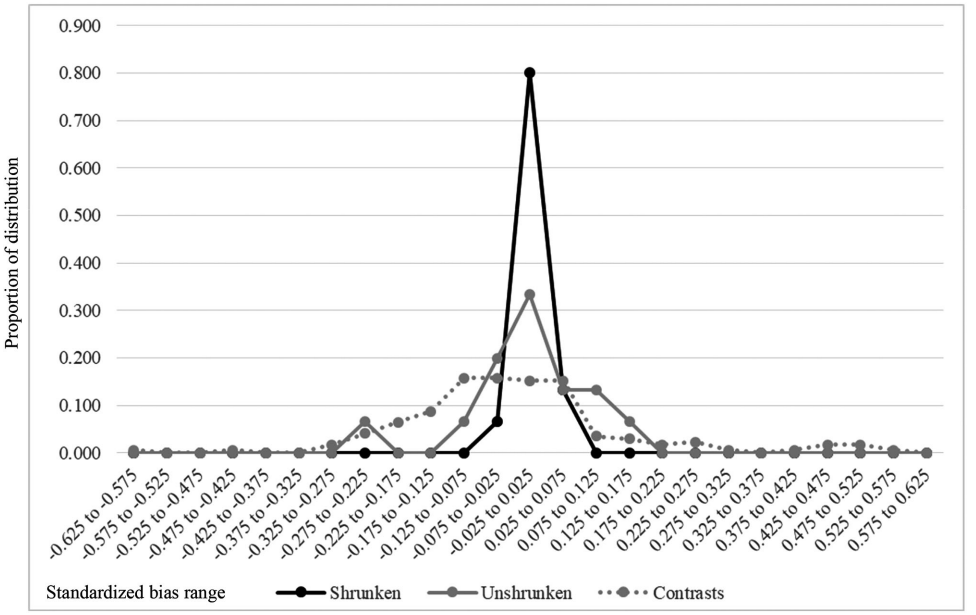
Figure 3. Distribution of Contrast-Level Bias Estimates.

Table 7. Frequency of contrast-level bias estimates by some design types.

		% With absolute bias>		
Subgroup description	Contrasts	0.10	0.20	0.30
All	171	44%	17%	8%
RD $N > 1{,}000$	124	43%	12%	4%
RD $N > 1{,}000$ and nonparametric	35	29%	9%	6%
RD $N > 1{,}100$	117	43%	10%	2%
RD $N > 1{,}100$ and nonparametric	33	24%	3%	0%

The Distribution of RD Bias Estimates across Contrasts

The preceding analyses focus on the average RD bias across all WSCs and on the average bias across all the contrasts within each WSC. However, the unit of analysis in this study was contrasts and not studies, and contrasts inevitably contain more random error. Figure 3 shows that the 171 contrasts have mean RD bias close to zero and that these contrast values are evenly distributed around this mean—the same pattern found at the study level. However, the distribution of RD bias estimates is now much larger, ranging from −0.578 to + 0.560. Indeed, Table 7 shows that just over 40 percent of the contrasts have absolute RD bias estimates above 0.10 and almost 20 percent are over 0.20. Even when the contrasts are limited to those analyzed nonparametrically and with RD sample sizes over 1,100, 24 percent of contrasts have bias over 0.10 and 3 percent have bias over 0.20. A simple linear regression analysis indicated that RD bias estimates were about 0.11 standard deviation units larger for studies with RD sample sizes under 1,100 and, within the



Notes: The shrunken estimates are from the Bayesian model without control variables and include the intercept. The unshrunken estimates are unweighted averages of the contrasts by study. The Contrasts line describes the unshrunken data by contrast. The unshrunken mean by study is −0.009. The mean by contrast is −0.024, which is lower because almost half of the contrasts come from Black, Galdo, and Smith (2007), which had a mean of −0.071.

Figure 4. Distribution of Bias Estimates by Type.

set of studies with larger sample sizes, the bias was about 0.04 standard deviation units larger when parametric RD methods were used. Both coefficient estimates had *P*-values < 0.01, allowing for the clustering of contrast estimates by study and including no other controls. Even so, the variation in estimated bias is unacceptably large at the contrast level. Analysts who pick out one contrast at random from the many that a study might generate will seldom recreate valid RD estimates. Fortunately, this is not how RD analysts typically behave as they seek to take advantage of current theoretical knowledge about which RD contrasts are technically superior to others.

SUMMARY AND DISCUSSION

The main results of this paper are summarized in Figure 4. It plots the median RD bias in small bins arranged from the highest to the lowest. The distributions are: (1) by contrast, (2) by study using unshrunken estimates, and (3) by study using shrunken estimates. The shrunken estimates are from the Bayesian analyses without covariates. Results from the frequentist analysis without covariates are very similar.

Three things stand out. First, in all analyses the mean, median, and modal bias estimates are close to zero. We can be confident, therefore, that, across this sample of 15 WSCs, RD impact estimates at the cutoff are close to those from RCTs at the cutoff. Second, in all study-level analyses, the distributions are symmetrical on each side of the mean of zero. Thus, there is no evidence that positive or negative RD bias is more prevalent. Third, the distribution of bias estimates varies by level of the analysis, being largest at the contrast level, next largest at the study level in

analyses without shrinkage where four of 15 RD bias estimates exceed 0.10, and being smallest with shrunken estimates at the study level where the distribution of bias estimates seems to be leptokurtic and none exceeds 0.070. The diffuse contrast-level distribution suggests that any one RD impact estimate from any one RD study is likely to be far from the RCT benchmark value. The study-level distribution of unshrunken estimates is less diffuse, but some RD bias estimates still exceed 0.10 and so imply that a modest fraction of study-level RD effects may be biased. However, the distribution of shrunken study-level estimates indicates little RD bias when the broader range of information is brought to bear that meta-analysis facilitates.

The question is: Which analyses should one trust most? The contrast-specific tests are the least trustworthy. First, they are subject to more unreliability than estimates of study-level averages, and this will inevitably increase the dispersion that is evident in Figure 4. Second, it would be odd RD practice for an analyst to select a single RD estimate at random from among all those plausible ones an analysis generated. More common is for researchers to generate multiple impact estimates that vary a set of realistic assumptions about bias or else for them to make an *a priori* case from theory and past empirical results in order to argue that some estimates are preferable to others—for example, those based on larger samples, or nonparametric analyses, or with certain ways of obtaining optimal bandwidth values.

The study-level unshrunken RD bias estimates reflect what an individual researcher would find in a single study examined in isolation from related empirical work. In four of the 15 cases in this synthesis, RD bias exceeded 0.10 standard deviation units, an often used but still arbitrary criterion for inferring whether to tolerate the level of bias obtained or whether not to do so. However, the studies in question were among the smallest WSCs, and theory indicates that RD estimates are less likely to be biased the larger the sample sizes are. Researchers of single RD studies cannot assume willy-nilly that the bias in their study will be close to zero, but they can increase the odds of this by a few well-known ameliorative procedures such as using larger samples, nonparametric tests, and careful bandwidth choices.

The study-level shrunken RD bias estimates take advantage of a crucial feature of meta-analysis: That information from all studies can be used to reduce the variance estimates that are needed in each study to compute RD bias. The mechanism for this is shrinkage—in this case, shrinkage to the mean of all studies. When this is done, all of the study-specific estimates are under 0.070 standard deviations. We place most faith in the shrunken estimates in Figure 4 because they use more of the data than the unshrunken study estimates.

Only a tight distribution of study-specific bias estimates allows policymakers to recommend RD in the confidence that its results will be minimally biased, or even unbiased, across all the research projects they support. It is worth remembering that in the Bayesian analysis there was a 97 percent probability of bias within 0.05 standard deviations of zero, and that in the frequentist analyses the 95 percent confidence interval around the average RD bias of essentially zero fell between 0.05 and −0.05. Such tight bounds should increase the confidence of all those who financially support causal research, should reduce any uncertainty among individual researchers hesitant to use RD in their own causal work, and should embolden teachers to recommend RD in their students' work. Each potential user group can be confident that RD works, not just on average, but also in individual RD studies that have larger samples and have been carefully conducted. Of course, legitimate debate is possible about how small RD/RCT differences must be to merit being called functionally equivalent. But we surmise that few researchers will see an average difference of 1/100th of a standard deviation, or study-level differences between 0.07 and −0.07, as substantively meaningful. So, we conclude that RD is generally internally valid at the treatment cutoff, just as statistical theory predicts it to be. Our second conclusion is that the distribution of study-specific

RD bias estimates is quite tight, implying that RD worked at the cutoff in the 15 ways it was implemented and in the 15 ways data from that design were analyzed. The theory undergirding basic RD can now be interpreted *as robustly internally valid in research practice*. The researchers who conducted these 15 WSC analyses all knew how to analyze RD data without succumbing to the pitfalls that can beset so assumption-dependent a method.[2] This should be comforting to three main stakeholder groups: those who make decisions about the causal methods to use in tests of hypotheses about policy, program, or project impacts; those who need to select a causal nonexperimental method in their own work; and those who teach causal methods and who can feel even more comfortable promoting RD where it is appropriate. The "slings and arrows of outrageous" implementation and analysis that potentially bedevil RD did not operate in debilitating ways across these 15 WSCs.

Of course, the RDs in this synthesis do not represent all the possible RDs to which policymakers, researchers, and instructors might want to generalize. We obviously cannot achieve the population of all possible studies comparing RD and RCT estimates in WSCs. Indeed, we cannot even be sure of having found all past relevant studies. We searched broadly and found 15 WSCs dealing with RD, of which eight were unpublished at the time (though two of those are now in press). Even so, we cannot be sure we detected all the relevant studies and that their results would be similar to those reported here. To address this file-drawer problem, meta-analysts often estimate how many studies would have to be overlooked to make the average obtained effect shrink to zero. The analog in the RD bias case is to ask how many studies with bias of size X are needed to move from essentially zero average bias to bias of size Y. In the WSC literature, Y is sometimes set to 0.10 standard deviations and, after shrinkage, the largest bias in any individual study was 0.07 in the analysis. If the missed studies had twice as much bias—0.14, and if they all had the same weight as the average included study, then we would need 35 undetected studies to make the average bias rise from zero to 0.10. We judge the odds of failing to locate so many studies as low, thus increasing our confidence in the robustness of the essentially zero average RD bias reported here.

Let us turn now to external validity. One-third of the WSCs were conducted as synthetic WSCs in which the RD data came from an RCT and not from a normal RD. However, more typical of general RD practice are the nine WSCs using the tie-breaker method that embeds an RCT along one segment of the assignment variable and where an RD structure occupies the rest of the assignment variable. The average bias with tie-breaker studies was no different from the other RD variants—essentially zero. Another external validity consideration is that one of the authors of this study (Cook) was a co-author on four of the WSCs (Kisbu-Sakarya, Cook, & Tang, in press; Shadish et al. 2011; Tang, Cook, & Kisbu-Sakarya, forthcoming; Wing & Cook, 2013). Many of the other authors of those four studies were or are also affiliated with Northwestern University. These researchers might be thought disposed to discovering that RD results are unbiased, given Northwestern's history in developing the design (Cook, 2008). However, RD bias estimates tended to be larger for these four studies than for the others, making it difficult to argue that long-time advocates of RD are especially likely to recreate similar RCT estimates.

Logically, there is no way to generalize beyond features common to all 15 of these WSCs. One common feature is that they were all conducted by researchers

[2] Common pitfalls include making unnecessarily strong assumptions about the relationship between the running variable and the outcome, ignoring the possibility of manipulation of the running variable, and allowing one's priors about the estimated impacts to affect decisions regarding functional form or bandwidth.

who were interested enough in methodology to want to test RD's internal validity. Do such researchers conduct better RDs than other researchers, given their likely superior technical skills and even their knowledge that their RD results are to be checked against an RCT's? The present study included some measures of RD quality, and the amount of bias was not related to any of them. While we cannot rule out that less sophisticated RD researchers might have obtained more average bias, this possibility has to be weighed against the fact that researchers can learn how to do technically better RDs. There is no shortage of papers and software offering documented guidance (e.g., Cattaneo, Titiunik, & Vazquez-Bare, 2017).

Another common feature of these 15 WSCs is that each included an RCT benchmark. Do the average bias results we reported here continue to hold when an RCT is conceptually impossible or logistically difficult—for example, when treatment noncompliance is likely to be large and impossible to control? There is no logical way to know this. However, it is worth remembering that the 15 studies we examined were quite heterogeneous in other dimensions and that this can be used inductively to argue that it increases the odds of finding no bias in other circumstances—for example, when an RCT is not possible. Also relevant is statistical theory, for it outlines the conditions under which RD estimates are unbiased asymptotically, and, to date, these conditions do not specify anything that is obviously akin to an RCT being impossible or logistically difficult. Clearly, these two arguments are oblique and do not "prove" the applicability of the present findings to situations where an RCT is not practical. The issue remains serious, making it clear that the present findings are more definitive with respect to internal rather than external validity, even though they are superior for external validity when compared to prior tests of RD's internal validity based on single datasets and their inevitably more limited populations of persons, settings, times, and cause and effect constructs.

In the WSC context, RD and RCT effects are estimated at essentially the same point. However, their estimation processes differ, principally in whether functional form assumptions are needed and how well supported they are. This estimation difference made little systematic difference to the estimates actually obtained across these 15 studies. It seems, then, that the complexities of estimation are real in RD theory but are less real in RD practice; whatever the 15 sets of researchers did was sufficient to rule out the potentially baneful influence of imperfect functional form estimation. Moreover, the fact that RD estimates were just as valid in the late 1990s, when advice about estimation was less sophisticated than today, suggests that recent advances in RD estimation methods may have only a small marginal benefit. That so little RD bias was obtained across such diverse WSCs should encourage the belief that RD is broadly effective in practice, even if it will never be universally effective across all possible applications.

Although we have empirically demonstrated the internal validity of basic RD and have partially extended its external validity, this does not mean that RD and RCT are equally informative. RD can only be used to estimate impacts close to the cutoff value(s) and, with sample size held constant, it has larger standard errors than an RCT. RD also requires more numerous and more opaque assumptions due to the absence of the crucial potential outcome slope representing what would have happened to treatment group members if they had not been treated. Moreover, the diagnostic tests for examining assumptions are also likely to be less definitive with RD. While both parametric and nonparametric traditions have evolved to minimize the internal validity threat in RD, they require correctly modelled functional forms or correctly specified bandwidths (e.g., Calonico, Cattaneo & Titiunik, 2014; Imbens & Kalyanaraman, 2009; Trochim, 1984). All this makes RD important as a complement to RCTs, as a method to use when RCTs are not possible or extremely difficult to implement—a proposition that we did not directly test since all WSCs of RD require an RCT. The evidence for lack of bias where only RD is feasible is indirect at

best—by induction from the heterogeneity of contexts in which RD's internal validity has now been demonstrated and theoretically from the proofs of RD's general lack of bias. The present study may seem brute empirical, but statistical theory underlies its prediction of zero RD bias and is also crucial for supporting extrapolation to contexts where an RD is possible but an RCT is not. The general tenor of the findings is clear, though. For all stakeholder groups who need to make decisions about causal methods worth promoting, the current meta-analysis provides them more information with which to advocate for basic RD.

DUNCAN D. CHAPLIN is a Senior Researcher at Mathematica Policy Research, 1100 1st Street, NE, Washington, DC 20002 (e-mail: DChaplin@Mathematica-mpr.com).

THOMAS D. COOK is a Research Professor at the Trachtenberg School of Public Policy at The George Washington University, 805 21st Street, NW, Washington DC 20052 (e-mail: t-cook@northwestern.edu).

JELENA ZUROVAC is a Senior Researcher at Mathematica Policy Research, 1100 1st Street, NE, Washington, DC 20002 (e-mail: JZurovac@mathematica.net).

JARED S. COOPERSMITH is a Statistician at Mathematica Policy Research, 1100 1st Street, NE, Washington, DC 20002 (e-mail: JCoopersmith@mathematica-mpr.com).

MARIEL M. FINUCANE is an Associate Director III at Mathematica Policy Research, 955 Massachusetts Avenue, Suite 801, Cambridge, MA 02139 (e-mail: MFinucane@mathematica.net).

LAUREN N. VOLLMER is a Statistician at Mathematica Policy Research, 955 Massachusetts Avenue, Suite 801, Cambridge, MA 02139 (e-mail: LVollmer@mathematica-mpr.com).

REBECCA E. MORRIS is a Graduate Student at the Milken Institute School of Public Health at The George Washington University, 950 New Hampshire Ave NW, Washington, DC 20052 (e-mail: remorris@gwu.edu).

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grants: DRL-1228866 and DGE-1544301. The authors would also like to thank Austin Nichols, Phil Gleason, Steve Glazerman, Heinrich Hock, Hanley Chiang, Larry Hedges, Peter Steiner, and reviewers from this journal for valuable advice given during the writing of this paper.

REFERENCES

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. Evaluation Review, 22, 207–244.

Ashworth, K. E., & Pullen, P. C. (2015). Comparing regression discontinuity and multivariate analyses of variance examining the effects of a vocabulary intervention for students at risk for reading disability. Learning Disability Quarterly, 38, 131–144.

Baker, S. G., & Lindeman, K. S. (2001). Rethinking historical controls. Biostatistics, 2, 383–396.

Barrera-Osorio, F., Filmer, D., & McIntyre, J. (2014). Randomized controlled trials and regression discontinuity estimations: An empirical comparison. Working paper. Cambridge, MA: Harvard University.

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. Journal of Experimental Criminology, 6, 191–208.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. Journal of Policy Analysis and Management, 31, 729–751.

Black, D., Galdo, J., & Smith, J. A. (2007). Evaluating the regression discontinuity design using experimental data. Unpublished manuscript. Chicago, IL: University of Chicago.

Buddelmeyer, H., & Skoufias, E. (2004). An evaluation of the performance of regression discontinuity design on PROGRESA. Bonn, Germany: IZA Institute of Labor Economics.

Calonico, S. M., Cattaneo, D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82, 2295–2326.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297–312.

Cattaneo, M. D., Titiunik, R., & Vazquez-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, 643–681.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics, 142, 636–654.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27, 724–750.

Cook, T. D., & Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. Annales d'Économie et de Statistique, Econometric Evaluation of Public Policies: Methods and Applications, 91/92, 127–150.

Cooper, K. G., Parkin, D. E., Garratt, A. M., & Grant, A. M. (1997). A randomized comparison of medical and hysteroscopic management in women consulting a gynaecologist for treatment of heavy menstrual loss. BJOG: An International Journal of Obstetrics & Gynaecology, 104, 1360–1366.

Dahabreh, I. J., Sheldrick, R. C., Paulus, J. K., Chung, M., Varvarigou, V., Jafri, H., ... Kitsios, G. D. (2012). Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. European Heart Journal, 33, 1893–1901.

Deeks, J. J., Dinnes, J., D'amico, R., Sowden, A. J., Sakarovitch, C., Song, F., ... Altman, D. G. (2003). Evaluating non-randomized intervention studies. Health Technology Assessment, 7, 1–173.

Efron, B., & Morris, C. N. (1977). Stein's paradox in statistics. Scientific American, 236, 119–127.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, Scotland: Oliver and Boyd.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness, 5, 189–211.

Glazerman, S., Levy, D. M., & Meyers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. The ANNALS of the American Academy of Political and Social Science, 589, 63–93.

Gleason, P. M., Resch, A. M., & Berk, J. A. (2012). Replicating experimental impact estimates using a regression discontinuity approach, No. 7461. Princeton, NJ: Mathematica Policy Research.

Goldberger, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations. Discussion paper 129–172. Madison, WI: University of Wisconsin, Madison, Institute for Research on Poverty.

Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S., & Larimer, C. W. (2009). Testing the accuracy of regression discontinuity analysis using experimental benchmarks. Political Analysis, 17, 400–417.

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. Handbook of Labor Economics, 3, 1865–2097.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1, 39–65.

Hyytinen, A., Meriläinen, J., Saarimaa, T., Toivanen, O., & Tukiainen, J. (2009). Does regression discontinuity design work? Evidence from random election outcomes. Helsinki, Finland: Government Institute for Economic Research Working Papers.

Imbens, G., & Kalyanaraman, K. (2009). Optimal bandwidth choice for the regression discontinuity estimator. NBER Working Paper Series No. 14726. Cambridge, MA: National Bureau of Economic Research.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142, 615–635.

Ioannidis J. (2005). Why most published research findings are false. PLoS Medicine, 2, 696–701.

Jung, H., & Pirog, M. A. (2014). What works best and when: Accounting for multiple sources of pure selection bias in program evaluations. Journal of Policy Analysis and Management, 33, 752–777.

Jung, H., & Pirog, M. A. (2017). Sample conditions under which bias in IV estimates can be signed. Journal of Policy Analysis and Management, 36, 909–932.

Kisbu-Sakarya, Y., Cook, T. D., & Tang, Y. (in press). Statistical power for the comparative regression discontinuity design with a non-equivalent comparison group. Psychological Methods. doi: 10.1037/met0000118

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. American Economic Review, 76, 604–620.

Lee, D. (2008). Randomized experiments from non-random selection in U.S. House elections. Journal of Econometrics, 142, 675–697.

Lord, F. M. (1960). Large sample covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 55, 307–321.

Moss, B. G., Yeaton, W. H., & Lloyd, J. E. (2014). Evaluating the effectiveness of developmental mathematics by embedding a randomized experiment within a regression discontinuity design. Educational Evaluation and Policy Analysis, 36, 170–185.

Nickerson, D. (2007). An evaluation of regression discontinuity techniques using experiments as a benchmark. PowerPoint poster presented at the annual meeting of the Society of Political Methodology, July 18–21, State College, PA.

Nickerson, D., Friedrichs, R. D., & King, D. C. (2006). Partisan mobilization campaigns in the field: Results from a statewide turnout experiment in Michigan. Political Research Quarterly, 59, 85–97.

Popper, K. (1934). The logic of scientific discovery. Heidelberg, Germany: Mohr Siebeck.

Popper, K. (1963). Conjectures and refutations: The growth of scientific knowledge. Abingdon, England: Routledge & Kegan Paul.

Rosenthal, R. (1979). The "File Drawer" problem and tolerance for null results. Psychological Bulletin, 86, 638–641.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. Journal of the American Statistical Association, 103, 1350–1353.

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

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. Psychological Methods, 16, 179–191.

Steiner, P., Cook, T. D., & Shadish, W. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. Journal of Educational and Behavioral Statistics, 36, 213–236.

Tang, Y., Cook, T. D., & Kisbu-Sakarya, Y. (forthcoming). Statistical power for the comparative regression discontinuity design with a pretest no-treatment control function: Theory and evidence from the national Head Start impact study. American Journal of Evaluation.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51, 309–317.

Trochim, W. M. K. (1984). Research design for program evaluation: The regression-discontinuity approach. Beverly Hills, CA: Sage.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? The Annals of the American Academy of Political and Social Science, 578, 50–70.

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of impact with education test scores as outcomes. Journal of Policy Analysis & Management, 26, 455–477.

Wing, V. C., & Cook, T. D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. Journal of Policy Analysis & Management, 32, 853–877.

APPENDIX

Details of the Nickerson (2007) Reanalysis

We reanalyzed the data from Nickerson (2007) in order to align the RCT–RD contrasts from that paper with the rest of the WSCs used in this meta-analysis. The data are described in Nickerson, Friedrichs, and King (2006) and essentially involve an RCT urging individuals aged 18 to 34 to turn out to vote. The RD involves adding to these cases data from people over the age of 34 and then testing whether the RCT and RD estimates agree around that age. The density of cases above 34 was much greater than below since there was no recruitment into a study above age 34. The population was essentially used.

Analysis of the RCT

To estimate a LATE with the RCT we restricted the RCT sample to ages 30 to 34, and estimated the following model using a linear regression:

$$V02g_i = \alpha + \beta_1 T_i + \beta_2 A_i + \beta_3 G_i + \beta_4 V02_{pi} + e_i\,, \tag{A.1}$$

where $V02g$ is the outcome (voting in 2002 general election), T is the treatment indicator receipt of a message urging voter turnout), A is age, G is gender, and $V02p$ is a dichotomous indicator for voting in the 2002 primary election. We chose the 2002 primary election as a proxy pretest for the 2002 general election because other available indicators were for the 1998 and 2000 elections. In the year 2000 there was a presidential election while in 2002 there was not, and 1998 was less preferable given the time difference. β_1 is interpreted as the treatment effect in this model. The estimated impact was a 0.063 increase in voter turnout, with a standard error of 0.012. Additional details are provided in Table A1. It includes the standard deviations used to standardize both the RD and RCT impact estimates.

Analysis of the Basic RD

Before estimating the RD model we first checked the data to see if there was evidence suggesting bias. A McCrary test for manipulation of the running variable was not possible because of the relatively limited density of cases between 30 and 34. However, we ran the RD analysis with variables representing turnout in prior elections as outcomes, including the 2002 primary election and the 2000 primary and general elections. The 1998 data were omitted from the falsification tests because they are missing for people who were below age 22 in 2002 since those individuals were not eligible to vote in 1998. This problem was less of an issue for the 2000 data. None of the estimates in these falsification tests was significantly different from zero, leading us to conclude that any effect on the 2002 general election outcome is

Table A1. Descriptive statistics from the randomized control trial.

Description	Treatment	Control
Observations	4,081	2,169
Adjusted mean outcome	0.540	0.480
Unadjusted standard deviation	0.498	0.500
Age range	30–34	30–34

Table A2. Falsification results from pre-intervention years.

Voting outcome	Impact at cutoff	*P*-value
2000 General	0.031 (0.020)	0.120
2000 Primary	0.019 (0.021)	0.352
2002 Primary	−0.020 (0.028)	0.481

Note: Standard errors in parentheses.

likely from the treatment, and not some other unmeasured cause. These tests were done using the same model used to estimate RD impacts except that the gender and 2002 primary election variables were omitted as controls. The falsification test results are shown in Table A2.

To estimate impacts on the 2002 general election outcome we used a local-linear regression with an Imbens & Kalyanaraman (2009) optimal bandwidth of 9.3 and a triangular kernel. This gave us a total of 21,522 observations (6,497 in the treatment group, below age 35; and 15,025 observations in the comparison group, age 35 and above). The model included age centered at the cutoff, the treatment indicator; age interacted with treatment status, an indicator for gender and an indicator for voting in the 2002 primary election.

$$V02g_i = \alpha + \beta_1 T_i + \beta_2 A_i + \beta_{at} T_i * A_i + \beta_3 G_i + \beta_4 V02p_i + e_i \quad \text{(A.2)}$$

The r-packages "rdd" and "rdrobust" were used for this analysis, and the estimated impact was 0.050 with a standard error of 0.018 and a *P*-value of 0.004. Bias is thus indicated as the difference between the unstandardized RCT estimate of 0.063 and the RD's estimate of 0.050. This difference was transformed into standard deviation units for inclusion in the meta-analysis.