

Annales d'Economie et de Statistique (in press)

Empirical Tests of the Validity of the Regression Discontinuity Design

Thomas D. Cook and Vivian C. Wong

Northwestern University

Abstract

This paper reviews the literature on whether regression-discontinuity studies reproduce the results of randomized experiments conducted on the same topic. After briefly reviewing the regression-discontinuity design and its history, we explicate the general conditions necessary for a strong test of correspondence in results when an experiment is used to validate any non-experimental method. In Economics, within-study comparisons of this kind are associated with LaLonde (1986), and we elaborate on how to do such studies better than twenty years ago. We identify three cases where regression discontinuity and experimental results with overlapping samples were explicitly contrasted. By criteria of both effect sizes and statistical significance patterns, we then show that each study produced similar results across the experiment and regression-discontinuity study. This correspondence is what theory predicts. But to achieve it in the complex social settings in which these within-study comparisons were carried out suggests that regression-discontinuity results may be generally robust.

The Regression Discontinuity Design

Theoretical research in the social sciences is often concerned with causation, understood as identifying the effects of causal agents that can be deliberately manipulated. The concern is particularly manifested in program evaluation where debates still occur about the roles that experiments and non-experiments should play in identifying the causal impacts of policies and programs. These debates are now closed in some sectors, like medicine, public health, mental health, criminal justice, the prevention sciences, and even some parts of social welfare. The randomized experiment now reigns supreme in these areas, institutionally supported through the privileged role it plays in graduate training, research funding and academic publishing. Of course, random assignment is not universally feasible; and in some areas, debate still persists about the merits of random and systematic treatment assignment. So social scientists will always need to know about causal methods other than the experiment that can produce unbiased causal inferences or, failing that, methods that generate bias so small it can sometimes be tolerated.

Several beliefs are widely used to justify the preference for random assignment. The major one is that, when perfectly implemented, such assignment creates comparison groups that do not differ in expectation other than for the consequences of treatment assignment. It is thus that selection is ruled out (Rubin, 1974; 1978). However, the regression-discontinuity design (RDD) also permits unbiased causal inference in theory. In this design, units are assigned to treatment based on a cutoff score on an assignment variable, often a score indicating a specific level of merit or need that then determines who is to receive or not receive some need- or merit-based resource. RDD is not limited to need and merit contexts. It has also been used with birthdates, like the date of entering formal schooling between January and December (Angrist &

Lavy, 1999; Ludwig, 2005), and even the order of applying for a particular resource. Although the design requires respecting the cutoff for treatment exposure, it is flexible as to what that assignment variable is so long as it forms at least an ordered continuum. For example, it is possible to index need by a single measure like household income (Lohr, 1972) or by a composite of many measures indicating physical hardship (Buddelmeyer & Skoufias, 2003) or by some index that predicts the likelihood of being out of work a long time (Black, Galdo & Smith, 2005). It is not important that the assignment variable has a clear meaning, or has high construct validity. Nor is it required that values on the assignment variable represent true scores since observed scores determine treatment exposure. The crucial need is for units to be assigned to one treatment if their assignment score falls below a given cutoff and for all units to be assigned to a different treatment if their scores are above it.

The theoretical key here is that the assignment process into treatment or control status is completely known (Goldberger, 1972a; 1972b). Since this is also an attribute of random assignment, Mosteller (1990) has suggested that RDD be considered as just another form of the experiment. It is not, though, in the sense that the efficiency of RDD is lower (Goldberger, 1972a; 1972b). Moreover, RDD is also not conceptually buttressed by as elegant a warrant as formal sampling theory provides for the experiment. Nor has there been as much experience with implementing RDD as with implementing experiments. Such experience is important if researchers are to identify the assumptions on which the method depends, if they are to be able to engineer how to test whether these assumptions hold and, most importantly, if they are to learn how to design research in order to prevent the assumptions from being violated in the first place. The history of RDD is manifestly more recent than the history of the experiment, and its implementation has not attracted as much attention—yet!

As interpreted by Cook (in press), the history of RDD began in Psychology with Thistlewaite & Campbell (1958). In Campbell & Stanley (1963), two rationales were offered for the design. The first was rooted in its similarity to the randomized experiment at points immediately around the cutoff score--and only there. The relevant intuition is that someone whose IQ is 140 is hardly different from someone whose IQ is 139, yet one gets the intervention for gifted students but the other does not. The selection difference between students with these two scores is likely to be almost entirely due to chance, to measurement error; and chance is the very mode of treatment assignment from which the experiment draws its interpretative power. The second rationale for RDD that Campbell offered was that the selection process into treatments can be very well indexed by the regression line of the assignment variable on outcome. Indeed, this value in the untreated part of the assignment variable provides the counterfactual against which a change in level or slope at the cutoff is assessed. The implication here is that the untreated slope functions like the untreated control group mean in an experiment. Campbell's work also specified the two major interpretative threats with RD—that bias can result if the functional form of the assignment on outcome is mis-specified or if treatment misallocations occur around the cutoff. Such misallocations are presumably more likely when the cut value is used to justify how scarce resources are distributed, and it was to explore this possibility that Campbell discussed the school leavers' exam in Ireland. There, a state exam was used for determining who could leave school, and scores just below the cutoff were systematically under-represented in the total distribution of scores, presumably because the examiners did not want to fail students so close to passing. This example points to social influences that modify allocation decisions in ways that do not completely respect the technical need to respect the intended cutoff.

Psychologists pursued Campbell's agenda for 40 years, learning how to analyze data so as to avoid mis-specifying functional forms, how to design studies so as to get independent checks on what the functional form in the treated part of the assignment would look like if there were no treatment there, how to identify and handle the fuzzy RDD that follows from treatment mis-allocations around the cut score, and how to determine and increase statistical power so as to minimize this disadvantage relative to the experiment. All these developments are summarized in Trochim (1984). Nonetheless, use of the design was quite sporadic in Psychology, and by 1995 it was essentially in abeyance. All those who had sought to make their reputations by pursuing it went on to study other things, and all that remained was its popularization in a few specialized texts on quasi-experimental design (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002).

In Statistics, the design never really caught on. It was mentioned as early as 1979 (Rubin, 1979). It was also the origin of several advances in non-parametric regression in order to deal more flexibly with the functional form assumptions on which the method so heavily depends (Spiegelman & Sacks, 1980). Also, Spiegelman (1976; 1977; 1979) conducted studies to find ways of dealing with the fuzzy discontinuity problem. The first formal proof in Statistics came later, and was due to Finkelstein & Robbins (1996a; 1996b). They called RDD the risk allocation design to emphasize how it could be used to assign patients to treatment based on some cutoff on a risk measure. It is clear that the authors had not then heard of RDD and thought they were offering a novel and general way of solving the selection bias problem. Indeed, the journal editors allowed them what was essentially three articles to make and illustrate their case, and even commissioned a special editorial from Mosteller (1996). Also in Statistics, Berk & Rauma (1983) generalized the design to deal with dichotomous rather than continuous outcomes (see also Berk & de Leeuw, 1999). But apart from that, nothing much happened in Statistics.

Why? It is impossible to be sure, but we suspect that to those formally trained statisticians who know about the design saw it as limited in its generality, as heavily tied to causal inference around a single cutoff value. Indeed, the generality issue emerges when comparing RDD to Rubin's version of assignment on the basis of a covariate. His model requires random assignment of the treatment at some points on the assignment variable but all-or-none assignment to one or the other treatment at the other points. The random assignment allows valid estimation of the functional form of the different treatment group regressions, and researchers can test how well these values correspond with adjacent values on the assignment variable where all-or-none assignment occurs. In this model, RDD is merely the special case of all-or-none allocation on one side or another of the cutoff without any random allocation at all. Moreover, the tie-breaking experiment (Boruch, 1973) that combines RDD with symmetrical random assignment in a region around the cutoff--thus formalizing what Campbell intuited gave RDD its inferential power anyway-- becomes in Rubin's formulation the special case where random assignment occurs at all assignment points within some range around the cutoff but is all-or-none outside this range. The point is that Rubin's formulation is more general than RDD, a virtue that scholars formally trained in Statistics are likely to appreciate. For them, identifying causal impacts when the selection process is completely known represents a trivial intellectual challenge; and dealing with functional forms and treatment mis-classification probably involve minor variants on well-known data analytic procedures that themselves command a quite minor place in the discipline. In this interpretation, it is not surprising that RDD is hardly visible in Statistics.

RDD's history in Economics is more complex, and only partially overlaps with its history in Statistics. The first formal proof of RDD came in two unpublished papers by the

econometrician Arthur Goldberger (1972a; 1972b). However, the proof was serendipitous and incidental to the main purpose of these papers. Indeed, Goldberger never knew he had provided proof of RDD as such, for he did not know about it at the time. In any event, the proof in Goldberger's own words, with our clarifications added in parentheses, boils down to the following: "Recall that z (a binary variable representing the treatment contrast at the cutoff) is completely determined by pretest score x (an obtained ability score). It cannot contain any information about x^* (true ability) that is not contained within x . Consequently, when we control on x as in the multiple regression, z has no explanatory power with respect to y (the outcome measured with error). More formally, the partial correlation of y and z controlling on x vanishes although the simple correlation of y and z is nonzero" (Goldberger, 1972,a; p. 16).

Goldberger's independent, serendipitous and unpublished discovery of RDD did not have much reverberation in Economics which, at the time, was pursuing causal method agendas more general than RDD. The search was for methods to handle selection in the many and varied circumstances where it occurs without a cut score. For a time, Heckman-type selection models and other uses of instrumental variables were the rage. But it became apparent from within-study comparisons like those of LaLonde (1986) and Fraker & Maynard (1987)--later followed by many other similar studies summarized in Glazerman, Levy & Meyers (2003)—that as Heckman-type models were used in actual research practice they failed to recreate the results of experiments that shared the same treatment group as the non-experiment. So the bloom went off the Heckman rose. Experience also began to accumulate that decisions about whether particular instrumental variable applications met the method's stringent requirements depended very heavily on social consensus rather than on independently verifiable technical criteria. This realization undermined faith in many applications of instrumental variables and so many

economists cast around for other causal methods, not being willing to rely on extrapolations from the complex substantive theories incorporated into general equilibrium models. So these younger economists turned to random assignment experiments, natural experiments, and RDD studies for the control each gave over exogeneity. After about 2000, a flurry of theoretical papers on RDD began to appear in *Economics* (e.g., Hahn, Todd & van der Klaauw, 2001; Lee, 2006), as did many applications of the design (e.g. Angrist & Lavy, 1999; van der Klaauw, 2002; Jacob & Lefgren, 2004a; Jacob & Lefgren 2004b; Ludwig & Miller, 2005; Gormley & Philips, 2005) and even a special edition of the *Journal of Econometrics* (in press). In addition, some American governmental agencies that commissioned evaluations began to specify that RDD should be used in preference to other methods if an experiment was not possible (IES, 2004). By about 2002, RDD seemed to have arrived in *Economics* after its birth elsewhere in 1958 and after its extremely low profile and independent discovery in that discipline in 1972.

Requirements for Validating RDD against Experiments in a Within-Study Comparison

It is one thing for a method to justify causal inference in some theory of method, and another thing for it to produce unbiased inferences in actual research practice. After all, one could glibly assert that complete knowledge of the selection process results in unbiased causal inference. Based on this, one might then try to measure the selection process as completely as possible and adjust study results for what one had learned about this process. While such practice respects theory, it is unlikely to generate the same results as an experiment. This is because selection processes are so complex in the real world that it is well-nigh impossible to conceptualize and measure them perfectly. For example, in job training some individuals sign up

because of court sanctions, others because of social worker suggestions, others because they are bored, others because their spouses or significant others persuade them to, others because their friends do so, others because they have heard of recently trained acquaintances who got stable jobs, and others out of sheer desperation. And, of course, there can be idiographic blends of all these different motivations. With RDD, the main practical tasks are to estimate functional form properly and to deal with any fuzziness there might be at the cut point, tasks that are difficult but still much easier than conceptualizing and measuring selection processes like those in job training.

But even so, with RDD we need to know not so much whether it can produce the same results as experiments, for theory predicts that it should when all the conditions of each method type are met. Rather, the issue is, “Do RDD and experimental studies produce the same casual estimates when they are implemented in the real world with all the warts that necessarily accompany the implementation of any one study, particularly multi-year, multi-site studies with complex interventions?” A related issue is whether the two types of studies produce similar standard errors, and consequently the same pattern of statistical significance. We know from Goldberger’s work that experiments should be about 2.75 times more efficient than RDD studies; and we know from other work that the actual degree of superiority--that is, the deviation from 2.75--depends on a number of factors outlined in Cappelleri, Darlington, and Trochim (1994) and in Shadish, Cook, and Campbell (2002). If standard errors are indeed larger in an RDD study than an experiment, in some circumstances this is bound to change the conclusions researchers draw based only on evidence from statistical significance patterns.

A methodology for answering questions like those proposed here already exists. It is the so-called within-study comparison study, of which LaLonde (1986) was the originator. He took

the causal estimate from an experiment and compared it to the estimate from a non-experiment that shared the same treatment group as the experiment but had a control group that was non-randomly formed. The intent was then to use OLS analyses and Heckman-type selection models or—in the work of LaLonde’s successors--propensity scores, to see if any of them could correct for the selection bias occurring in the non-experiment. The non-experiments in question were all studies of two non-equivalent groups that had a posttest observation, pretest demographic information and sometimes one or more pretest scores on the same scale as the outcome. None of the non-experiments in the review of this literature by Glazerman et al. (2003) was an RDD study. Yet it is easy to imagine the same kind of within-study comparison being used to contrast experimental results with those from RDD, though as we see below some modifications have to be made to suit the particularities of RDD.

As Smith and Todd (2005) pointed out, some of the past within-study comparisons were technically flawed. It is indeed difficult to conduct a good study of this type, and so we explicate below seven criteria that, if met, improve the interpretation of a within-study comparison.

1). A within-study comparison has to demonstrate variation in the types of method being contrasted—one comparison group has to be constructed via a random assignment mechanism and the other by whatever systematic mechanism is under test. In the RDD case, this should be assignment via a cutoff score. This difference in assignment mechanisms constitutes the independent variable of interest in a within-study comparison. It is the putative causal agent, as it were.

2). The two assignment mechanisms cannot be correlated with other factors that are related to the study outcome. As Smith & Todd (2005) showed with the LaLonde-type job training studies, it is not a good idea to construct the non-equivalent comparison group from

some national data source like the Current Population Survey. This confounds the experimental and non-experimental study with how the outcome was measured, where it was measured, and when it was measured, as well as with geographic setting. Identifying the consequences of different assignment mechanisms requires that no correlates of the outcome should be confounded with the assignment mechanism. In the RDD case, such a confounding would arise if, for example, units one side of the cutoff were measured differently from units on the other side.

3). A quality within-study comparison also has to demonstrate that the randomized experiment deserves its status as a causal gold standard. This implies that a correct randomization procedure has been used, that it has been correctly implemented, that there is not unhappy randomization, that there is not differential attrition, and that there are no treatment crossovers to vitiate the SUTVA conditional assumption (Rubin, 1990). It would be futile indeed to compare non-experimental results to experimental ones that failed to meet these assumptions, for the experiment could not then function as the validity criterion for the non-experimental study results. In practice, all experiments deviate from most of these assumptions in various ways, increasing standard errors in many cases and even producing some bias in others, even if only to a minor degree. This makes the “gold standard” rhetoric a little stretched, although the argument is strong that no other causal method is as well fitted to serving this function as the well-conducted experiment.

4). It is also important that the non-experiment be a good example of its type. This is a somewhat more opaque requirement and its specifics obviously depend on which type of non-experiment is under discussion. In the RDD case, there is more clarity than with most other non-experimental designs. We know the three major things that need to be examined: how the

functional form issue is handled; how the analyst deals with treatment mis-allocations around the cutoff; and how the analyst deals with the lesser statistical power of the RDD study. Needed for functional form are descriptions of the slopes each side of the cutoff. Where they are not linear and parallel, adjustments for non-linearities have to be made within a parametric perspective or else semi- or non-parametric methods must be used. Also, efforts need to be made to see how estimates vary as a function of proximity to the cutoff. Greater faith is warranted in estimates generated from a narrower bandwidth around the cutoff, for this is not as sensitive to functional form mis-specification as broader bandwidths. But narrower bandwidths create a trade-off with statistical power by decreasing sample size. So this trade-off also has to be taken into account in evaluating how well an RDD study has been done. Concern about the technical quality of the non-experiment is central because we want to identify whether its causal results closely approximate those from an experiment when the design was executed as well as the state of the art allows, as opposed to it being executed in whatever fashion an analyst happens to have accomplished. While this last evaluates how well a particular application of a given design approximates the results of an experiment, it does not evaluate the potential of a design to reproduce the results of an experiment.

5). An experiment and non-experiment should estimate the same causal quantity. It makes little sense to compute the average effect size for an experiment and to compare it to the local average treatment effect from an RDD study that is estimated at a different point on the assignment variable than the average in the experiment. Should the relationship between the assignment variable and outcome be non-linear, there is no reason to expect correspondent effect sizes. This puts a special premium on being able to identify those experiments where the experiment takes place among cases symmetrically distributed around the cut score. Only then is

the experiment's average treatment effect strictly comparable with the local average treatment effect in an RDD study. Of course, comparison at different points on the assignment variable is not meaningless, though it can be mischievous. In particular, it is important for assessing the robustness concern of this paper. Will similar results be achieved when the causal entities being estimated are not quite identical, as well as in the technically more correct context when they are?

6). A within-study comparison should be explicit about the criteria it uses for inferring correspondence between experimental and non-experimental results. Identical estimates are not to be expected. Even close replications of the same randomized experiment will not result in identical posttest sample means and variances. Assuming statistical power of .80, the same pattern of statistical significance will result in only 68% of comparisons. That is, the probability of two significant findings across experiments is $.80 \times .80$, and the probability of two non-significant ones is $.20 \times .20$. Comparisons of significance test patterns are important in the RDD context, not because so many scholars use them for decision-making, but because RDD studies are known to be less efficient than experiments. All things being equal, then, differences in statistical significance patterns can be expected across an experiment and the RDD study yoked to it, with the latter study producing more no difference findings. More meaningful than statistical tests are focused tests of the difference between casual estimates from an experiment and RDD study. But these are rare in the literature we review. Instead, we will report estimates for each method type singly, using our own judgment to make decisions about comparability and inviting readers to judge for themselves.

7). The data analyst should perform the non-experimental analyses before learning the results of the experimental ones. This is to prevent the analyst deliberately or inadvertently

skewing the many decision points in any analysis so as to increase the likelihood of recreating the already known experimental results. Such demand characteristics are common and powerful in science (Rosenthal, 1966). None of the studies reviewed provides evidence about the temporal sequence of method comparisons, and so we have little to say about this point except to note that it is desirable for future research in this area. However, it is worth noting that RDD imposes more discipline on the analyst than most other types of non-experimental study, given how transparent is the assignment process. This makes it more difficult to play around with many selection models before deciding on the one that produces closest correspondence to the experiment. Some playing around is still possible, particularly in the choice of functional form specifications and other covariates. But the extent of this is almost certainly less than with other non-experiments.

Having presented our criteria for evaluating quality within-study comparisons, we proceed with the following three goals: 1). to identify all RDD and experimental comparison studies that were simultaneously conducted on the same topic; 2). To assess the degree of correspondence by design type according to criteria of similarity in both their causal estimates and statistical significance patterns; and 3). to use the seven criteria above to begin the process of explaining whatever degree of causal outcome correspondence was achieved. We found three such studies that compared results from an experiment with those from an RDD, and discuss them in depth below.

Aiken, West, Schwalm, Carroll and Hsiung (1998)

Aiken et al. (1998) examined how students enrolled in a college remedial writing class performed relative to students who did not take the course. The RDD took advantage of a school policy that assigned students to a remedial writing course on the basis of their ACT or SAT scores, depending on which of these they had taken during high school. The treatment group (N=99) consisted of students who scored below the particular test-specific cutoff score and had consented to participating in the evaluation study prior to the start of the school year. These students took the college remedial writing class in the fall semester, followed by the standard freshman composition course in the spring semester. The comparison group (N=119) consisted of students who had scored above the cutoff, and were randomly selected from fall or spring semester sections of the standard freshman composition, stratified across times and days. All students took the writing assessments as part of in-class testing exercises at the beginning and end of the semester. The RDD analysis was an ANCOVA using the assignment variable, the cutoff score, and the pretest score as covariates. Separate analyses were conducted on each of the two assignment variables, SAT and ACT scores. The authors tested the robustness of the functional form assumption by adding a quadratic selection term and an interaction between the selection term and binary treatment predictor, and found no appreciable change in effect size estimates.

The randomized experiment involved taking a sample of students whose “admission test scores...fell within a fixed range just below the normal cutoff scores” (Aiken et al., 1998, pg. 212). These students were asked to volunteer to be in an experiment that would assign them either to taking the remedial course or going straight into standard English writing classes. No section of standard freshman composition enrolled more than one control participant to ensure that class progress would be unaffected by the presence of excess non-remediation students.

Treatment students (N=39) assigned to the remedial course were administered the writing assessment at the beginning and end of fall semester, and again when they completed the standard freshman composition class at the end of spring semester. Control students (N=69) were administered writing assessment outside of class at the beginning of fall semester, and again as part of an in-class examination in their standard freshman composition course. Controlling for pretest scores from the beginning of fall semester, Aiken et al. used ANCOVA to compare posttest writing assessment scores for remediation versus non-remediation students at the end of fall and spring semesters.

Despite the challenge of implementing a study at a major public university with multiple campuses, a vast bureaucratic structure, and nearly 50,000 enrolled students, Aiken et al. (1998) met four of our seven criteria for a strong test of design types. The study varied in whether assignment was at random or via cutoff. It generally succeeded in holding most irrelevancies constant, with both designs drawing their samples from the same pool of university students, having participants undergo experiences together at the same institution, and assessing students under similar conditions. Dependent variables for both studies were also the same: performance on the TSWE, and on an essay exam scored by at least two independent trained raters. We did observe, however, one extraneous confound that might be correlated with the outcome and discuss this below. Randomization appeared to have worked, with no significant differences on pretest scores for the TSWE and the writing sample and no differential attrition. Finally, the authors were explicit in their criteria for correspondence, examining both the pattern of effect size as well as the pattern of statistical significance.

The experiment and RDD, however, diverged in at least two ways. First, the randomized experiment was not the tie-breaker design usually advocated in the RDD literature, with its

choice of units symmetrically distributed around the cutoff (Boruch, 1975; Campbell, 1984). Instead, the authors estimated their average treatment effect in the experiment *at a point just below* where they estimated the local average treatment effect in their RDD study. Also, the RDD sample had to take the course to which their ACT or SAT scores assigned them, whereas students in the randomized experiment could refuse the randomization invitation and still stay at the university. Variation in the sample selection procedure confounds within-study results because discordant findings could be driven by sample differences rather than the assignment mechanism itself.

The RDD appeared adequate, with no reported instances of fuzzy discontinuity. However, the authors were fortunate to have obtained robust results at all given the small sample size of their RDD. Separate analyses were conducted on the SAT and ACT assignment variables, with 35 and 46 students in the SAT and ACT treatment groups and 87 and 72 students in the SAT and ACT comparison groups. Sensitivity tests indicated that the relationship between the assignment and outcome variables was parallel and linear. Had it not been so, the authors would have had difficulty generating results through either parametric or non-parametric means. The small sample size would have also limited the power of their RDD, making detection of any significant effect difficult. It was likely the inclusion of pretest measures that enhanced the power of the RDD to a level where significant effects were detectable. Overall, we believe that Aiken et al. (1998) presented a fair – but imperfect – test of design types, and that they were lucky to have generated stable RDD results at all, much less achieve the level of correspondence that they report.

Table 1 summarizes results for the experimental and RDD studies. The randomized experiment produced a statistically significant standardized effect size of .59 on the TWSE. The

RDD produced a reliable standardized effect of .49 when the assignment variable was ACT and a non-reliable .32 when SAT was used. On the writing task, all effect sizes were non-significant—in the experiment it was .16 and in the RDD it was .22 for the ACT assignment variable and .02 for the SAT. The authors noted that the *pattern of effect sizes* was the same for the randomized experiment and the RDD – the three largest effects were for TWSE and the three smallest were for the writing test, irrespective of design type. Correspondence in *similar patterns of statistical significance* for experimental and RDD results were found in half the TSWE results, and in all of the writing assessment results. Taken together, Aiken et al. (1998) concluded that their experiment and RDD study produced generally comparable results, with the bases for this claim being quite clear for the ACT assignment variable, less so for SAT.

Buddelmeyer and Skoufias (2003)

The second experiment/RDD contrast was by Buddelmeyer and Skoufias (2003). They reanalyzed data from PROGRESA, a major Mexican program aimed at alleviating current poverty through monetary and in-kind benefits, as well as reducing future levels of poverty through investments in education, nutrition, and health. The authors used the fact that Mexican villages were randomly assigned to PROGRESA, but that families within the experimental villages were then assigned into treatment conditions based on their score on a scale of material resources. For the experimental and RDD studies, the authors examined whether PROGRESA improved school attendance and reduced labor force participation among girls and boys between the ages of 12 and 16. In total, 3301 boys and 2941 girls were assigned to the treatment condition in both the RDD and experimental studies, 1563 boys and 1378 girls were assigned to the

comparison group for the RDD study, and 1952 boys and 1863 girls were assigned to the control group for the experimental study. One round of pre-intervention data and two rounds of follow-up data were analyzed for this study.

The assignment variable in the RDD was a composite of detailed census information on all households and individuals living in communities covered by the program. Those scoring below the threshold were considered “poor” and eligible for participation in the program while those scoring above the cutoff were considered “not poor” and did not receive government support. Because treatment communities were located in seven different states in Mexico, separate composite variables and thresholds had to be calculated for each locality to account for geographic variations in material poverty. The selection method led to approximately 52% of households in the evaluation sample to be classified as eligible for program benefits.

The RDD data were analyzed using non-parametric procedures and required information from only the following sources: the outcome measures, the poverty scale scores, and the seven cutoffs that varied from region to region. To rule out spurious discontinuities due to the assignment and outcome variables being non-linearly related, the authors used one-sided kernel regressions to estimate the unconditional means of the outcome measures. Their main analyses presented treatment estimates for a variety of kernel functions (rectangular, biweight, triangular, quartic, Epanechnikov, and Gaussian), and subsequent analyses showed estimates at different bandwidths (50, 75, and 100). Despite claims by PROGRESA central administration officials that assignment into the program was not based on a “purely mechanical approach in the sense that selected households were reclassified from one category to another based on an additional set of filters such as age, [and] feedback from local authorities...”, the authors found little evidence that non-eligible households were actually reassigned into the treatment condition.

Thus, they characterized the PROGRESA selection process as a “sharp” rather than “fuzzy” RDD.

The experimental study compared program-eligible families in villages randomly assigned to PROGRESA with similarly eligible families in the control villages. The analyses used pooled data from all survey rounds, but included controls for round of survey and for household, individual, and geographic characteristics. The authors report that randomization worked at the locality level, with no significant differences in means on key variables. However, significant differences in means were detected when individual level data was examined, suggesting that assignment was not entirely random because observed characteristics in the pre-program round significantly predicted assignment into the treatment condition. As a result, the authors report estimates from cross-sectional analyses that compared post-program treatment and control group means, as well as estimates from difference-in-difference analyses.

In assessing the comparability of the experimental and non-experimental designs, we note variation in assignment mechanisms, and attempts to reduce confounds correlated with study outcomes. Measurement seems to have been the same in both the experimental and control villages and also for the treatment and non-equivalent comparison groups within the experimental, and hence also RDD, villages. As discussed above, pretest results indicated problems with the experiment, so the authors used adjusted experimental estimates to address randomization concerns. They found little evidence of fuzziness in the RDD data, and employed non-parametric procedures to avoid assumptions about the functional form. Both the experimental and RDD studies had large sample sizes, though the RDD had fewer comparison groups members than the experimental. The slightly smaller sample size in the RDD likely had

minimal consequences on the study's power given that the experimental sample had just over 10,000 cases while the RDD had almost 10,000.

The main area of discrepancy between the two design types lies in the localities of estimated treatment effects. The experimental study estimated average treatment effects (ATE) while the RDD study estimated local average treatment effects (LATE) for a subgroup of observations near the cutoff. The authors attempted to address this concern by presenting experimental estimates for a *restricted sample of households* that had assignment scores within the same RDD bandwidth. A comparison of the cross-sectional estimates with the restricted sample estimates revealed evidence of heterogeneity in program impacts across households with different profiling scores, thus prompting the authors to use experimental results for the restricted sample as their benchmark estimates. To be sure, using restricted sample of the experimental group is an ad hoc and generally not recommended method for ensuring that the same causal quantities are estimated. This is because the RDD analyses was conducted within experimental villages where the comparison groups were more materially advantaged on average than were the randomly formed control groups in the non-experimental villages. Thus, the average effect for the experimental design inevitably fell below the local average treatment effect at the cutoff for the RDD. While the restricted sample likely reduced differences in average treatment effects, some discrepancy in causal quantities estimated surely remained.

In Table 2, we summarize Buddelmeyer and Skoufias' findings. We present the authors' preferred experimental estimates – the “restricted sample” results – for the benchmark results, and estimates using six kernel functions for the RDD results. Overall, the RDD estimates confirmed the experimental findings that PROGRESA had little to no impact on boys' and girls' work activities in both rounds of follow-up. Close correspondence was also achieved for boys'

and girls' school attendance in the second round of follow-up, but not for the first round, with RDD showing no significant program effects on school attendance and the experimental study showing positive and significant effects. Looking across all outcomes for both boys and girls, the rate of agreement for RDD and experimental estimates by *pattern of statistical significance* was over .80 in six of eight possible rounds, and if the first round of follow-up for school attendance is discounted, then close correspondence was achieved in every round. The authors write that if “one were to put aside, for the moment, the discrepancies observed in [the first round of follow-up for school attendance], the performance of the RDD appears to be remarkably good.”

The authors examined why the RDD produced such divergent results from the experimental design for the first follow-up round of school attendance. They found that increasing the bandwidth of the non-parametric estimators provided a partial explanation, with improved performance of the RDD estimator for girl's school attendance, but not for boys. The authors then hypothesized that spillover effects may have contaminated the comparison group in the RDD study. Spillover might occur if noneligible households in treatment communities began altering their behavior by enrolling their children in school due to peer effects, expectation for benefit receipt, or any other reason. To test their hypothesis, the authors generated estimates using a variety of comparison groups (including those just above the cutoff in the experimental control villages), and found that spillover was a plausible explanation for the poor performance of the RDD estimates during the first round of follow-up, but that spillover effects had disappeared by the second round. They concluded that “it is the comparison group rather than the method itself that is primarily responsible for the poor performance of the RDD in [the first follow-up round].”

Black, Galdo and Smith (2005)

The third direct comparison of experiment/RDD results is the most methodologically advanced and in our estimation, the strongest test of design types. Black, Galdo and Smith (2005) reanalyzed data from a job training program in Kentucky that assigned potential exhaustees of unemployment insurance with mandatory reemployment services such as job-training and job-search workshops as a requirement for receiving benefits.

The RDD analysis took advantage of the fact that individuals were assigned to job training based on a single score derived from a 140-item test predicting the likelihood of long-term unemployment. The assignment scale used five years of administrative data on claimants' past earnings, schooling, past job characteristics, prior unemployment receipt, prior welfare receipt, industry and occupation controls, and local economic and labor market conditions. For each local employment office in each week, new claimants were ranked by their assigned scores. Reemployment services were given to those with the highest scores, followed by those with the next highest scores until the slots for each office each week were filled. When offices reached their maximum capacity for claimants to receive employment services, and if there were two or more claimants with the same profiling scores, then random number generators were used to assign claimants into the treatment condition. Thus, only claimants with marginal profiling scores – the one at which the capacity constraint was reached in a given week and in a given local office – were randomly assigned into experimental treatment and control groups. This sampling procedure resulted in a true tie-breaking experiment and ensured that the RDD causal estimate was at the same average point on the assignment variable as the experiment, creating a more interpretable contrast of the two design types.

Following Rosenbaum's (1987) suggestion, the authors also used two alternative comparison groups in the RDD to better identify program impacts. The "selection bias from above" estimates consisted of the RDD treatment group (group D in Figure 1) and the RDD comparison (group A) and experimental control (group C) groups, while the "selection bias from below" consisted of the RDD (group D) and experimental (group B) treatment groups and the RDD comparison group (group A). Within each RDD sample, the authors matched treated and untreated individuals conditional on week and local office. The potential number of regression discontinuity groups, from both above and below, was bounded by the total number of marginal profiling groups. Thus, the authors considered only RDD groups that had corresponding tie breaking groups in a given office for a given week, ensuring comparable samples between the experiment and RD. In all, 1222 and 742 claimants were in the experimental treated and control groups, and 46,313 and 8,631 claimants formed the RDD treatment and comparison groups. The large sample size for the RDD suggests that the authors were appropriately concerned about creating enough power in the RDD to detect treatment effects.

Both parametric and non-parametric procedures were used to generate RDD estimates for three outcomes – weeks receiving unemployment insurance (UI) benefits, amount of UI benefits received, and annual earnings. The parametric models included individual controls such as age, sex, race/ethnicity, and other variables not used in the profiling score. Cross-validation methods were employed to choose the appropriate order of the assignment variable. Since the authors assumed that assignment and outcome variables were not linearly related, they also presented three sets of non-parametric results: a simple local Wald estimator that took mean differences on raw outcome variables for "neighbors" at both sides of the discontinuity frontier, a smooth version of the Wald estimator that used multivariate kernel regressions to control for within-

group effects, and a one-sided unconditional kernel estimator (Hahn, Todd, and van der Klaauw, 2001) to address concerns about fuzziness at the discontinuity. The cutoff score varied by office and by week, and so treatment impacts were identified by first re-centering profiling scores to zero, and then by pulling the data together, weighting by the proportion of treatment units within each group. This design feature also had the advantage of allowing the authors to identify treatment effects over a frontier of discontinuity points, instead of at a single threshold.

The experimental impacts were estimated by differencing the mean outcomes of treated and untreated individuals within each site and week that also included a RDD treatment and comparison group. Effects were then summed and weighted by the proportion of treated units in each marginal profiling group. The authors explain that the “experimental estimates can be thought as a weighted average of the estimates from many small randomized experiments” (pg. 32).

To assess comparability of results, the authors computed non-experimental bias by taking the difference between RDD and experimental impacts, and then by testing for significant differences in means. Because of space constraints, we present RDD estimates in Table 3 for only observations that were closest to the cutoff ($\pm .05$ from discontinuity), though the authors also presented estimates at $\pm .10$ from the cutoff, $\pm .15$ from the cutoff, and for the full sample. The authors found no significant differences between experimental and RDD results for any of the parametric estimates. Additional results showed that the parametric models yielded less biased results when the sample was restricted to observations closest to the cutoff. Bias for the outcomes “weeks receiving UI benefits”, “amount of UI benefits received”, and “annual earnings” increased from .04, \$4, and \$-70 at $\pm .05$ from the cutoff to -.85, \$-161, and \$-351 at $\pm .15$ from the cutoff. However, the level of bias varied with the sample used and the outcome

of interest. Estimates for the “selection bias from below” sample were generally less stable than results for the “selection bias from above” sample, and annual earnings estimates tended to exhibit large biases and were sensitive to the sample used.

For the non-parametric results, two of the three estimators performed well while the third method, the simple Wald estimator, yielded biased estimates for all outcomes and for both RDD samples. Only .67 estimates using the simple Wald estimator (at +/- .05 from the cutoff) were not significantly different from their experimental benchmarks, prompting the authors to write that the poor performance “highlights the importance of using pre-treatment covariates in the estimation of the conditional mean counterfactual for the outcome of interest” (pg. 36). In general, the multivariate “smooth” kernel estimator and the one-sided local linear kernel “HTV” regressions achieved closer correspondences, with no significant differences found in .83 instances using the smooth and HTV estimators. RDD estimates tended to diverge from their experimental benchmark for the outcome “amount of UI benefits received”, and for the “selection bias from below” sample in the multivariate kernel regressions.

Similar to the parametric estimates, the degree of correspondence was greater the more the RDD study was restricted to cases around the cutoff, again reflecting the non-linear function relating assignment to outcome. Thus, we see statistical theory about RDD guiding the analyses and elucidating one condition where an experiment and RDD study can generate comparable causal conclusions irrespective of functional form—viz., at a very local average treatment effect around the cutoff, given a non-linear functional form.

Finally, it is worth reviewing the features that make Black, Galdo, and Smith’s (2005) study the strongest test of RDD we have found. First, the randomized experiment and the RDD studies appear to be implemented well and analyzed correctly, with p-values for the test of

difference in means between the experiment and control groups showing no difference on almost all covariates in the experiment and significant differences between groups in the RDD. The use of three different types of non-parametric estimators addressed concerns about functional form in the RD, and the one-sided local linear kernel estimator is perhaps the best known method for handling fuzzy discontinuity if one wants to avoid assumptions about the relationship between assignment and outcome variables (Hahn, Todd, and van der Klaauw, 2001). Second, the authors went great lengths to ensure comparability between the RDD and experimental samples, using data only when there were claimants in both the RDD and experimental conditions in the same office during the same weeks. Third, the tie-breaking experiment imbedded within the RDD design allowed for the same treatment effects to be identified in both designs. Finally, the authors explicitly tested for statistical differences between experimental and RDD means using bootstrapped standard errors. In all, we regard Black, Galdo and Smith's study as being highly successful in meeting most of our criteria for a fair test of design types.

Discussion

What are we to make of these three within-study comparisons of experimental and RDD estimates? Generally, the studies were good examples of a within-study comparison, albeit not perfect ones. All created the intended contrast between assignment at random and by a cutoff score; all succeeded in ruling out the most obvious temporal and spatial third variable confounds with design type; all involved experiments that were apparently well conducted, though the pretest mean difference in PROGRESA at the individual-level (but not at the village-level) raises some minor doubts that are partially laid to rest by the difference-in-differences estimates. And

we have taken pains to show the degree of experiment/RDD concordance in results in terms of both causal estimates and statistical significance patterns.

Less reassuring is that none of the studies took pains to render analysts of the RDD data blind to the results of the experiment. This is a very important requirement, the more so when sample sizes are small or when the functional forms are not linear and parallel on each side of the cutoff. The first case is exemplified by Aiken et al. (1998). Even though pretest writing scores were in the RDD model, the sample sizes of 122 and 118 per cutoff score is quite small for estimating functional form well. And their ANCOVA analysis makes strong assumptions about linearity that are not easily tested under the conditions of the study. We wonder whether the analysis would have stopped where it did had there been no experimental estimate against which to validate the RDD results obtained? And had there been further analyses, possibly non-parametric ones, would they have created alternative estimates that added to the uncertainty about the size or presence of a causal impact? Consider Black et al. (2005) next. The non-linearity they found rightly induced them to try non-parametric analyses. Using multiple samples, smoothers, and bandwidths, they generated many RDD effect sizes that are quite variable (see Table 3). While there is some guidance as to which smoothers and bandwidths are to be preferred, it is far from perfect and one wonders to what extent the researchers' judgments as to which estimates are to be preferred were partly conditioned by knowledge of the experimental results.

Also not totally reassuring is that two of the studies failed to estimate the treatment effect at the same point on the assignment variable, thus confounding method type and the causal entity estimated. Black et al. (2005) were commendably careful to estimate identical quantities, conducting their randomized experiment symmetrically around the cutoff. Aiken et al. (1998),

however, did not. They estimated their experimental average treatment effect at a point just below the cutoff that defines the local average treatment effect in RDD. In Buddelmeyer & Skoufias (2003), the discrepancy in causal entities estimated would have been even greater, given that families selected by need in treatment villages were inevitably more materially disadvantaged than families at the cutoff for PROGRESA. But the analysts recognized this potential confound and so reconfigured their experiment to give an adjusted estimate closer to the cutoff. But even so it could not have been at the cutoff. Every treated person in the treated villages had to score below the cutoff to be in the reconfigured experiment, as did all those in the control villages who would have been treated had they lived in a treatment village. Thus, each contrast group in the experiment had to fall below the cutoff. We can surmise from the results that this particular confound played at best a minor role in biasing RDD results. Otherwise, the results would not have been so similar by design type. This correspondence of results despite a clear confound is surely because the average and local average treatment effects were tested at very close though not identical points. Achieving different estimates for each case would have required an abrupt non-linearity in the causal function beginning at a point immediately below the cutoff. By shrinking the difference in where the two effects were estimated, the chance of spurious effects in the RDD studies of Aiken et al. and Buddelmeyer & Skoufias were reduced, but not necessarily eliminated.

Good within-study comparisons should also involve contrasting a technically good experiment with a technically good example of the type of non-experiment under analysis—in this case, RDD. Aiken et al. (1998) were in some senses lucky to obtain the correspondences they did, given their small sample sizes, use of a simple ANCOVA, and failure to examine cases of fuzzy allocation in detail. Buddelmeyer & Skoufias' (2003) analysis was not as sophisticated

(nor as clear) as Black et al.'s (2005) with respect to the range and appropriateness of smoothers and bandwidths selected. Instead, the authors' chose to focus their analyses on examining six different types of kernel estimates, which the non-parametric literature has already shown to be a secondary concern after bandwidth selection (Pagan & Ullah, 1999). They found few instances of mis-allocation in the RDD data, but because Mexican officials' insisted that families were assigned into treatment conditions beyond what scores indicated, additional steps to ward against fuzziness might have been warranted.

As stated earlier, we regard Black, Galdo, and Smith's (2005) comparison to be the highest quality and most technically sophisticated of the three studies presented. However, their findings emphasized a distinct limitation in the state of the art in RDD studies without functional forms that are parallel and linear or lacking a pretest measure of the outcome that assesses pre-intervention functional form. Current theory for choosing smoothers and bandwidths is not very specific, and this leads to the need for multiple sensitivity tests that do not always add light. Indeed, Black et al. (2005)'s findings varied by the particular choices made, and an analyst who generated fewer options might have selected those most distant from the experimental estimate and thus have concluded that the RDD study was biased. Of course, some analytic practices seem a priori superior and so are to be preferred today, like the Hahn, Todd and van der Klauuw's (2001) one-sided local linear estimator. Other procedures are more dubious and Black et al. may only have used them in exploratory fashion, as with the simple Wald estimator that led to the most discrepant results (see Table 3). The point is that it will not be easy to recommend the best analysis without stronger theory and evidence about which non-parametric procedures to prefer when pretests and parallel and linear regressions are absent. One might argue that the resulting uncertainty about which causal estimate or estimates to pick in RDD is well deserved,

given the real uncertainty about how to smooth functional forms and select bandwidths. But it is a source of uncertainty one would not experience with an experiment, and it makes the results of any one RDD study less credible than it might otherwise be.

The limitations of these studies aside, there was still considerable correspondence between the experimental and RDD results. It was not perfect, of course, even within the probabilistic parameters that condition the level of correspondence to be expected. In Buddlemeyer & Skoufias (2003), the experiment and RDD study differed in the first year school outcomes for both boys and girls, and in Black et al. (2005), the broader bandwidths and the simple Wald estimator produced less concordance between the experiment and RDD. Buddelmeyer and Skoufias argued that the discordance was due to bandwidth choice and spillover effects in the comparison, and one can easily construct rationales for why broader bandwidths and simple Wald estimators performed worse than the other estimators in Black et al. (2005).

Even so, we are impressed by the level of correspondence achieved despite 1). the limitations of the studies as within-study comparisons, 2). each study being multi-year, multi-outcome, multi-context and involving highly complex treatments, and 3). the studies taking place in the largest state university in the USA, in Mexican villages, and in job training centers throughout Kentucky. There was ample opportunity in each case for many slings and arrows of outrageous implementation to have influenced sources of both error and bias that could have accumulated in opaque ways to create larger discrepancies between the experimental and RDD results. But this was not the case. The RDD estimates proved to be robust across all these factors. While it might not be theoretically exciting to show that experimental and RDD estimates are concordant, given that each should produce unbiased estimates, it is distinctly promising to show

that each of them produces similar estimates in each of three cases involving complex studies in which many things can go awry for reasons of inadequate analyst theoretical knowledge, the logistical realities of all complex empirical work, and the state of the art limitations of RDD methodology. The three studies examined here suggest the robust ability of well- (but not perfectly-) analyzed RDD studies to recreate the results of experiments.

This is a promising finding that will achieve even greater solidity with more within-study comparisons that observe the seven criteria for such studies that are outlined in this paper. It will be easy to achieve analyst blinding, the estimation of identical causal quantities, the use of pretest measures of functional form, and greater discipline in the choice of smoothers and bandwidths for those cases where RDD regressions are not linear and parallel. The difficulty comes with the limited state of the art in non-parametric regression, the major impediment to comparing good experiments with good RDD studies. We hope that one side effect of this paper will be to get empirically oriented scholars to go out and do the better comparisons of experimental and RDD studies that we think are needed to be sure that RDD studies will generate mostly unbiased causal inferences in the hurly-burly of actual research life where all designs are implemented with some degree of imprecision and where there is no experiment to act as an external validation criterion. This paper adds an empirical component to the existing theoretical warrant for RDD; and within-study comparisons that are only slightly better will cement this empirical component and create from it an additional warrant for using RDD in cause-probing social research, including evaluations of public-sector policies and programs.

References

- Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, 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(4), 207-244.
- Angrist, J. D., & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*, 114, 533-576.
- Berk, R. A., & de Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association*, 94(448), 1045-1052.
- Berk, R. A., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, 78(381), 21-27.
- Black, D., Galdo, J., & Smith, J. C. (2005). Evaluating the regression discontinuity design using experimental data. *Working paper*.
- Boruch, R. (1973). Regression-discontinuity designs revisited, *Northwestern University*. Evanston, IL: Northwestern University.
- Boruch, R. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods and Research*, 4, 31-53.
- Buddelmeyer, H., & Skoufias, E. (2003). *An evaluation of the performance of regression discontinuity design on PROGRESA*. Bonn, Germany: IZA.
- Campbell, D. T. (1984). Forward. In W. M. K. Trochim (Ed.), *Research design for program evaluation* (pp. 15-43). Beverly Hills, CA: Sage Publications.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.
- Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review*, 18, 141-152.
- Cook, T. D. (in press). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.
- Finkelstein, M., Levin, B., & Robbins, H. (1996a). Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. *Journal of Public Health*, 86(5), 691-695.
- Finkelstein, M., Levin, B., & Robbins, H. (1996b). Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *Journal of Public Health*, 86(5), 696-705.
- Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, 22(2), 194-227.
- Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy*, 589, 63-93.
- Goldberger, A. S. (1972a). Selection bias in evaluating treatment effects: Some formal illustrations. Madison, WI: Institute for Research on Poverty.
- Goldberger, A. S. (1972b). Selection Bias in Evaluating Treatment Effects: The case of interaction. Madison, WI: Institute for Research on Poverty.
- Goldberger, A. S. (2006). Personal communication. In T. D. Cook (Ed.). Evanston, IL.

- Gormley, W. T., & Phillips, D. (2005). The Effects of Universal Pre-K in Oklahoma: Research Highlights and Policy Implications. *The Policy Studies Journal* 33(1), 65-81.
- Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201-209.
- Institute of Education Sciences (2004). Reading comprehension and reading scale-up research grants request for applications. Washington, DC: Department of Education.
- Jacob, B., & Lefgren, L. (2004a). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, 39(1), 50-79.
- Jacob, B., & Lefgren, L. (2004b). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, LXXXVI(1), 226-244.
- LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *The American Economic Review*, 76(4), 604-620.
- Lohr, B. W. (1972). *An historical view of the research on the factors related to the utilization of health services*. Bureau for Health Services Research and Evaluation, Social and Economic Analysis Division, Rockville, MD.
- Ludwig, J., & Miller, D. L. (2005). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design: National Bureau of Economic Research.
- Mosteller, F. (1990). Improving research methodology: An overview. In L. Sechrest, E. Perrin & J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (pp. 221-230). Rockville, MD: U.S. Public Health Service, Agency for Health Care Policy and Research.
- Mosteller, F. (1996). Editorial: The promise of risk-based allocation trials in assessing new treatments. *Journal of Public Health*, 86(5), 622-623.
- Pagan, A., & Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge, UK: Cambridge University Press.
- Rosenbaum, P. (1987). The Role of a Second Control Group in an Observational Study. *Statistical Science*, 2(3), 292-316.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34-58.
- Rubin, D. B. (1990). Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference*, 25, 279-292.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
- Smith, J. C., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators. *Journal of Econometrics*, 125, 305-353.
- Spiegelman, C. (1976). *Two methods of analyzing a nonrandomized experiment "adaptive" regression and a solution to Reiersol's problem*. Unpublished dissertation, Northwestern University, Evanston, IL.
- Spiegelman, C. (1977). A technique for analyzing a pretest-posttest nonrandomized field experiment. *Statistics Report M435*.

- Spiegelman, C. (1979). On estimating the slope of a straight line when both variables are subject to error. *The Annals of Statistics*, 7(1), 201-206.
- Spiegelman, C., & Sacks, J. (1980). Consistent window estimation in nonparametric regression. *The Annals of Statistics*, 8(2), 240-246.
- Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51, 309-317.

Table 1: Comparison of experimental and RDD estimates from Aiken et al. (1998)

Outcome	Experimental estimates		RDD estimates		Correspondence	
	Effect size	SE	SAT (ES)	ACT (ES)	Rate of agreement by pattern of effect size direction	Rate of agreement by pattern of statistical significance
TSWE	0.59*		0.32	0.49*	1.00	0.50
Writing assessment	0.16		0.02	0.22	1.00	1.00
N Treatment	39		35	46		
N Control	69		87	72		

* Significant treatment effect at .05 level.

RDD estimates in **bold** indicate lack of correspondence between RDD and experimental estimates in terms of statistical significance.

Table 2: Comparison of experimental and RDD estimates from Buddelmeyer & Skoufias (2003)

Outcome	Experimental estimate	RDD estimates						Correspondence Rate of agreement by pattern of statistical significance ~
		Uniform	Biweight	Epanechnik	Triangular	Quartic	Guassian	
<i>Boys</i>								
School attendance 1	0.07*	.02	.01	.01	.01	.01	.01	0.00
School attendance 2	0.10*	.05	.07*	.07*	.07*	.07*	.06*	0.83
Work participation 1	-0.01	.01	.00	.00	.00	.00	.01	1.00
Work participation 2	-0.04	-.03	-.03	-.03	-.03	-.03	-.03	1.00
<i>Girls</i>								
School attendance 1	0.08*	.04	.04	.04	.04	.04	.05*	0.17
School attendance 2	0.10*	.08*	.11*	.10*	.11*	.11*	.08*	1.00
Work participation 1	0.00	.01	.00	.00	.00	.00	-.01	1.00
Work participation 2	-0.03	-.02	-.03	-.03	-.03	-.03	-.03	1.00
N Treatment								
<i>Boys</i>	3301	3301						
<i>Girls</i>	2941	2941						
N Control								
<i>Boys</i>	1952	1563						
<i>Girls</i>	1863	1378						

* Significant treatment effect at .05 level.

~ Rate of agreement by pattern of statistical significance was calculated by looking at instances of correspondence in the pattern of statistical significance between experimental estimates and for each RDD estimates using the six different kernel functions. (biweight, epanechnik, triangular, quartic, and guassian).

RDD estimates in **bold** indicate lack of correspondence between RDD and experimental estimates in terms of statistical significance.

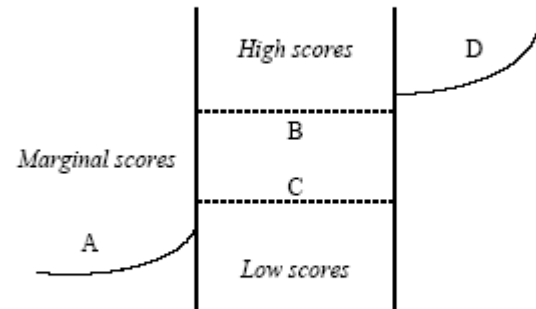
Table 3: Comparison of experimental and RDD estimates from Black, Galdo, and Smith (2005)

Outcome	Exp	RDD estimates ⁺				Correspondence			
		RDD estimate at +/- .05 from discontinuity ^{#+}				Rate of agreement by statistically significant diff in EXP & RDD means			
	Estimate	Para	Smooth	HTV	Wald	Para	Smooth	HTV	Wald
<i>Selection bias from above</i>									
Weeks receiving UI	-1.85	-1.81	-2.1	-2.04	-0.97				
Amount of UI benefits	-7.3	-3.3	-12.3	190.7	414.7				
Annual earnings	1338	1268	745	2659	1051				
<i>Selection bias from below</i>						1.00	.83	.83	.67
Weeks receiving UI	-1.92	-2.12	-2.65	-2.43	-1.75				
Amount of UI benefits	-22.9	-242.9	49.1	-18.9	200.1				
Annual earnings	1376	160	943	55	830				
N Treatment	1222	46,313							
N Control	742	8,631							

[#] We present only the RD estimates that were closest to the cutoff (+/- .05 from discontinuity), though Black, Galdo, and Smith (2005) also present estimates at +/-10 from the discontinuity, +/- from the discontinuity, and for all units. RD estimates in **bold** indicate significant mean differences between experimental and RD estimates using bootstrapped standard errors for the test statistic. Rate of agreement was calculated by looking at the number of significant differences between experimental and RD estimates at +/- .05 from the discontinuity.

⁺ The parametric estimator is referred to as “Para.” The three non-parametric estimators are referred to as the following: “smooth” for the smoothed version of the Wald estimator that uses multivariate kernel regressions to control for within-group effects; “HTV” for one-sided unconditional kernel estimator recommended by Hahn, Todd, and van der Klaauw (2001); and “Wald” for the simple Wald estimator;

Figure 1: Discontinuous Treatment Assignment
Kentucky Working and Reemployment Services



Notes:
B=Experimental treated group
C=Experimental control group
D=Non-experimental treated group
A=Non-experimental comparison group