PROJECT MUSE®

The Regression Discontinuity Design and the Social Corruption of Quantitative Indicators

Vivian C. Wong, Coady Wing

Observational Studies, Volume 3, Issue 2, 2017, pp. 183-209 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2017.0006>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/793390/summary>

The Regression Discontinuity Design and the Social Corruption of Quantitative Indicators

Vivian C. Wong

*Curry School of Education
University of Virginia
Charlottesville, VA*

vcw2n@virginia.edu

Coady Wing

*School of Public and Environmental Affairs
Indiana University
Bloomington, IN*

cwing@indiana.edu

Abstract

Thistlethwaite and Campbell (1960) (TC) introduced the Regression Discontinuity Design (RDD) as a strategy for learning about the causal effects of interventions in 1960. Their introduction highlights the most important strengths and weaknesses of the RDD. The main points of the original paper have held up well to more formal scrutiny. However, TC did not address “manipulation of assignment scores” as an important validity threat to the design. The insight that manipulation is a central validity threat is the most important conceptual advance in the methodological literature since its introduction. Although most modern RDD analyses include density tests for assessing manipulation, results are most convincing when diagnostic probes are used to address specific, plausible threats to validity. In this paper, we examine validity threats to two common RD designs used to evaluate the effects of No Child Left Behind and state pre-kindergarten programs.

Keywords: Regression Discontinuity Design, Imprecise Control, Manipulation of Assignment Scores, Threats to Validity, Quasi-Experiments, Donald T Campbell

1. Introduction

Thistlethwaite and Campbell (1960) introduced the Regression Discontinuity Design (RDD) as a strategy for learning about the causal effects of interventions in 1960. The basic features of the RDD are present in the initial article. However, TC underestimated the broad applicability of the design for applied work. They suggested that RDDs were apt to be rare in the real world. But these days, the truth seems quite the opposite. Organizations often adopt discontinuous assignment rules for practical or ethical reasons and researchers are able to put these rules to use to study a range of topics. A 2010 review of the RDD in economics, for example, listed 60 applications of RDD, including work on education, criminal justice, health, environmental policy, and politics (Lee and Lemieux, 2010). It seems clear that TC has had an important influence on quasi-experimental research.

TC used the bulk of their initial paper to discuss the strengths and weaknesses of the RDD. They made intuitive and heuristic arguments, which have held up well under more formal analysis developed in recent years. A lot of the conceptual work in the paper revolves

around the regression function linking average outcomes with assignment scores. They use a diagram to focus the discussion on the behavior of the regression function in the area near the cutoff score, and argue that a discontinuity in the level of the regression is strong evidence of a causal effect. They also make the careful point that a change in the slope of the regression is not convincing evidence of a causal effect because it seems more plausible that the regression function is non-linear. In other parts of the paper, TC interpret the RDD as an approximation to a randomized experiment among people with assignment scores near the cutoff.

The modern literature on RDD clarifies the underlying structure of both of these arguments. Hahn et al. (2001) showed that the validity of the RDD does not rest on the claim that the regression function takes a particular functional form, or on the assumption that information from below the cutoff can be extrapolated beyond the cutoff. The key assumption is that – at least near the cutoff – the regression function is a smooth one that does not have any naturally occurring breaks. Likewise, Lee (2008) presented the RDD in a local random assignment framework that revealed the sense in which RDD treatment effects could be viewed as a randomized experiment near the cutoff. His approach suggests that many of the key assumptions in an RDD will be satisfied as long as subjects are unable to exert precise control over their own assignment scores. McCrary (2008) proposed strategies for partially testing implications of imprecise control, which now play an important role in applied RDD studies.

The possibility that assignment scores could be manipulated seems to be one of the main insights of the modern RDD literature that was not part of the original TC article. The other features of the modern RDD literature – non-parametric regressions, methods for model selection, strategies for computing standard errors – are technical innovations that, while important and non-trivial, are logical extensions of TC’s original article. Assumptions about manipulation and precise control, however, seem to be a new thing. In this paper, we discuss the idea of manipulation and sorting in RDD studies, the logic behind the most common statistical tests used to detect manipulation, and the way that these tests fit into the overall “Campbellian” approach to quasi-experimental empirical work.

2. The Corruption of Social Indicators

A typical RDD takes advantage of a known and discontinuous assignment mechanism for the purposes of causal inference. The design often produces convincing evidence. But problems may arise when the researcher is not the only one who discovers the assignment rule. In a lot of cases, the assignment mechanism is public knowledge well before people are assigned to treatments. If the rule allocates a resource that is in high demand, it may create a strong incentive for people to manipulate their assignment scores to secure access to the resource. For example, suppose students are assigned to treatments according to their score on a qualifying exam. If they know the cutoff score in advance, they can study hard in an effort to make the grade. That kind of manipulation is not apt to be a problem. But an RDD study may be in trouble if the exam graders also have preferences over treatment assignments, and are willing to add or subtract points in ways that advantage particular students who have characteristics related to their outcomes. The ability of the grader to

exert fine control over the assignment score in non-random ways means that manipulation is possible, and may introduce selection bias near the cutoff.

Although the possibility that assignment scores may be manipulated seems like an obvious possibility, the implications for the RDD are subtle. Lee (2008) and McCrary (2008) showed that the RDD is robust to manipulation of the assignment scores as long as people can only manipulate scores with some degree of error. In fact, two conditions are required for manipulation to occur. First, there must be some mechanism for precise control over assignment scores. Second, the decision to manipulate must be correlated with other factors that also affect the outcome. When these two things occur, then the treatment effect estimates from an RDD may be biased.

3. Precise Control and Manipulation

An easy way to understand the idea of precise and imprecise manipulation is to represent the assignment score as a simple measurement error model. Let $A_i = e_i + u_i$ be a person's realized assignment score. In the model, e_i is a deterministic component that may be affected by factors like a person's effort and ability. In contrast, u_i is a stochastic error component that represents some set of conditions that are beyond the person's direct control. The basic point in Lee (2008) is that the internal validity of RDD depends on the assumption that individual realizations of u_i come from a distribution with a continuous density. In that case, there is a distribution of realized values of A_i even conditional on the value of the effort driven factors contained in e_i . When realized assignment scores are generated by both effort and error in this way, Lee says that people have imprecise control over the assignment process. And – more or less – he argues that RDD has high internal validity in such cases. On the other hand, people have precise control over the assignment process whenever assignment scores are determined purely by effort and do not depend on an error component with a smooth density. In other words, the internal validity of the RDD is a claim that realized values of the treatment variable are a discontinuous function of the assignment score, and that the assignment score is a continuous function of the error component.

The assumption that people do not have precise control over their assignment scores has some important implications (Lee, 2008). First, imprecise control over assignment scores implies that the distribution of observed and unobserved pretreatment covariates should be the same just below and just above the assignment cutoff. Second, the discontinuity in the conditional mean function linking outcomes to assignment scores represents an average causal effect of treatment exposure. In other words, when people do not have precise control over assignment scores, covariates should be balanced at the cutoff and the discontinuity in the conditional mean outcome function represents an average causal effect. As DiNardo and Lee (2010) point out, the important thing here is that these two statements (covariate balance and causal identification at the cutoff) are not “primitive assumptions” about the design. Instead, they are consequences of the treatment assignment procedure created by an RDD when people have imprecise control over their own assignment scores.

Hahn et al. (2001) (HTV) also showed that you could attach a causal interpretation to an RDD analysis. However, their justification is based on the “primitive assumption” that the relevant conditional mean functions satisfy smoothness assumptions. HTV did not offer much guidance to researchers trying to decide if smoothness was a reasonable

assumption in any particular study. When researchers invoke a smooth conditional mean assumption, they are making a statement about the data generating process that is ancillary to the research design. The assumption may or may not be true in the application at hand, and there is nothing about the design that makes smoothness more or less plausible. In contrast, Lee’s point is that local smoothness is implied whenever it is reasonable to believe that people have imprecise control over the assignment scores (Lee, 2008). Researchers can marshal qualitative arguments – institutional details – that help justify the assumption of imprecise control. And they can also use covariate data to test some of the implications of imprecise control, such as covariate balance at the cutoff and smoothness of the density of the assignment variable at the cutoff. Under Lee’s conceptualization of RDD, a critic cannot simply say that he believes the conditional mean function might not be smooth in the absence of the treatment effect. Instead, the critic must supply a plausible explanation for how a person might precisely manipulate their own assignment score. The virtue of the imprecise control assumption is that it encourages debates about concrete threats to validity and how to avoid them and test for them.

4. Testing The Imprecise Control Assumption

Lee (2008) and McCrary (2008) propose testable implications of the assumption that people have imprecise control over their own assignment scores. Lee (2008) showed that when assignment scores depend on both a deterministic and a random component, the distribution of pretreatment covariates should be the same just above and just below the cutoff. A simple way to test this implication of the imprecise control assumption is to perform the standard RDD analysis using pretreatment covariates as dependent variables. If the analysis finds that the mean of one or more covariates differs substantially above and below the cutoff, then the evidence might be used to reject the null hypothesis that people exerted no more than imprecise control over their assignment scores.

McCrary (2008) pursued a different implication. He reasoned that precise manipulation would distort the distribution of assignment scores by creating a discontinuity in the density of observed assignment scores at the cutoff. To implement McCrary’s test, the researcher constructs a histogram based on bins that radiate out from the cutoff score so that none of the bins overlap the cutoff. The researcher then uses local linear kernel regression to smooth out the histogram separately on both sides of the cutoff. If the analysis finds that the height of the density function is discontinuous at the cutoff, then the McCrary test would reject the null hypothesis of that people have only imprecise control over their assignment scores.

Combined, these tests suggest an approach that is distinctly Campbellian in its method for establishing causal inference. First, generate specific and plausible threats to validity that are based on substantive knowledge of how the assignment process was implemented. Second, hypothesize data patterns that should emerge if such validity threats are realized. Third, examine the empirical data for evidence of such threats. In the RDD, balance and density tests at the cutoff are essential tools for ruling out alternative explanations, and/or for diagnosing problems that may undermine the design.

Most modern RD analysis include some version of covariate balance and density tests for assessing the validity of the RD. This is a good thing. But we want to stress that these tests are most convincing when they form part of a coherent discussion of specific

threats to validity. Reflexively presenting tests about covariate balance and the density of the assignment variable as part of a list of “standard procedures” is probably not the best practice. It is important to ask whether the most likely threats to validity are apt to be captured by balancing tests and density plot. In a given application, the researcher may not observe all relevant covariates related to the outcome for balance tests. Or (more rarely, perhaps) observed discontinuities in estimated density functions at the cutoff may have an ambiguous interpretation. For example, a failed density test may indicate evidence of participants sorting into specific treatment conditions, but it may also uncover naturally occurring “heaping” on specific values of the assignment score. If these data generating processes do not induce correlations in third variables between the assignment variable and outcome, then discontinuities in the density function are not likely to generate selection bias.

In the remainder of the commentary, we consider two examples in which “sorting” behaviors around an RDD cutoff are suspected. The first is analysis of No Child Left Behind (NCLB) accountability rules for determining whether schools make Adequate Yearly Progress (AYP). The second involves the use of age cutoffs for evaluating the effects of state pre-kindergarten (pre-K) programs. In both cases, we adopt a Campbellian approach for ruling out threats to validity: Describe conditions under which sorting around an RD cutoff is suspected; formulate hypotheses about the data pattern you would expect under such threats, and examine data to assess the empirical evidence for these threats.

5. Empirical Example 1: Accountability Rules for Making Adequate Yearly Progress

One of the public goals of NCLB was to make all students “proficient” by state specific standards by 2014. To make the goal measurable, states established Adequate Yearly Progress (AYP) targets. Schools were supposed to meet proficiency thresholds on five indices: 1) the percentage of students proficient in reading according to the state reading assessment; 2) the percentage of students proficient in mathematics according to state mathematics assessment; 3) the percentage of students who participate in state reading assessments; 4) the percentage of students who participate in state mathematics assessments; and 5) at least one other academic indicator at each school level (elementary, middle, and high school). All schools held accountable had to meet AYP targets for the school as a whole and for any student subgroup that exceeds a state-set minimum number of students. In this analysis, we focus on rules that hold the student with disability (SWD) subgroup accountable for proficiency standards.

Although SWD subgroups were required to meet annual minimum percent proficiency thresholds in reading and math, schools that failed to meet these targets were allowed to apply an exemption under ancillary rules that lowered the effective cutoff for making AYP. States had discretion in determining exemption rules. But many states adopted two specific types of rules: confidence interval exemptions and safe harbor exemptions. Confidence interval exemptions work by attaching a “plus or minus” band around the percent proficiency target. To make AYP, a school needed only to achieve a target score that was equivalent to the lower bound of the confidence interval. The safe harbor rule works by examining a school’s performance in earlier years and allowing an exemption

based on adequate improvement. For example, if a school fails to meet the standard AYP, it would qualify for an exemption if 10 percent fewer students were not proficient this year as compared to the prior year.

Since AYP rules and exemptions are systematically and uniformly applied to all schools in the state and are public knowledge, the RDD has been applied as a method for evaluating the effectiveness of accountability policies. For example, Gill et al. (2008) used RDD to evaluate the effects of schools entering improvement status under No Child Left Behind, and Ahn and Vigdor (2014) applied a similar approach to examine the impacts of NCLB sanction rules on public North Carolina schools.¹

In earlier research, Figlio and Getzler (2006) found that schools reclassify and change the composition of students who are labeled as SWDs to meet accountability requirements. This suggests that RDDs based on proficiency thresholds may fail to meet the imprecise control assumption that justifies the standard analysis. Schools may be able to adjust their proficiency thresholds for the SWD subgroup if they are able to take advantage of the confidence interval and safe harbor exemption rules. For example, schools may be able to exercise precise control over the number of SWD students in their student body. Reducing the number of students in the SWD subgroup increases the width of the confidence interval and reduces the effective proficiency threshold. Likewise, schools may use the safe harbor rule by altering the composition of students in the testing group the following year so that a fewer percentage of students are considered not proficient. Schools with proficiency scores close to their cutoff may be more motivated to manipulate their threshold requirements than schools that are unlikely to make AYP for the year.

To assess the plausibility of these RD threats for SWD subgroup, we examined 2007-08 AYP data from Pennsylvania schools from grades three through eight. The sample includes only public schools that were held accountable under federal NCLB policy and have an eligible SWD subgroup (schools with 40 SWDs or more). To focus the discussion on schools subject to the discontinuous assignment procedures, we excluded data on schools that made AYP via the appeals process or missed AYP because of the participation and attendance requirements from our sample. In total, 1,035 public elementary and middle schools are included in the analysis sample, where 385 of these schools missed AYP in 2007-08; the remaining 645 schools made the cutoff.

To examine the presence of a discontinuity in the density of schools at their effective proficiency thresholds, we collapsed the multiple assignment rules for which the SWD subgroup could fail AYP into a single centered assignment variable and cutoff, reducing a high-dimensional assignment procedure into a single assignment mechanism.² We did this by first centering each school's reading and math-specific proficiency score around its effective cutoff. The effective cutoff for the school depends on the state proficiency threshold for the subject, as well as exemption rules such as the confidence interval or safe harbor rules that lower the proficiency requirements for each school. Once we obtained the centered assignment scores for reading and math for each school's SWD subgroup, we chose the minimum centered value (for either reading or math) as the school's assignment score.

1. In Ahn and Vigdors (2014) RD analysis, the results of the density tests and covariate balance checks indicated that sorting around the effective proficiency cutoff was not an issue in their sample.

2. Wong et al. (2012) refer to this as the "centering approach" for addressing multiple assignment variables and cutoffs in an RDD.

If schools took actions to undermine the annual accountability process, you might expect to observe data in which surprisingly few schools had proficiency rates immediately below the proficiency threshold. And you might expect to see a surprisingly large number of schools with proficiency rates at or immediately above the threshold. The density test for manipulation provides a statistical method to put this theory to the test. To conduct the density test, we constructed a histogram such that no bins included observations on the right and left sides of the cutoff. McCrary (2008) proposes using a bin size that is equal to $\hat{b} = 2\hat{\sigma}n^{-1/2}$ where $\hat{b}$ is the estimated bin size, $\hat{\sigma}$ is the sample standard deviation of the percent proficiency scores for SWDs, and n is the number of schools. We applied local linear kernel regression to smooth out the histogram separately on both sides of the cutoff, where the midpoint of each bin is treated as the independent variable and the number of observations falling in each bin is treated as the outcome. We chose a bandwidth using the method described in Imbens and Kalyanaraman (2012), and assessed the discontinuity at the cutoff by estimating the log difference in the height of the density at the cutoff. To see the estimator of the discontinuity in the density more concretely, let $\hat{f}^+$ and $\hat{f}^-$ be estimates of the height of the density function just above and below the assignment cutoff. We follow McCrary (2008) and estimate the discontinuity in the density as $\hat{\theta} = \ln \hat{f}^+ - \ln \hat{f}^-$. Under the null hypothesis that schools do not have precise control over the fraction of students who meet the proficiency goals of the state assessments, we expect $\hat{\theta} = 0$ and we form a simple bootstrap t-test to reject the null of imprecise control.

Figure 1 presents a frequency plot of centered proficiency scores for schools in Pennsylvania. The plot shows a clear dip in the number of schools immediately before the cutoff, followed by a sharp increase just over the cutoff. The dip suggests that schools scoring below the threshold manipulated their proficiency scores to make the AYP threshold; the spike suggests that there are many more schools at or above the cutoff than should be there if no sorting had occurred. In fact, there were almost 15 times as many schools at and above the threshold as there were below it. Table 1 presents results from a formal test of the discontinuity in the density of schools at cutoff. As shown in row 1, the discontinuity in log difference for the adjusted cutoff was large (1.51) and significantly different from zero (t-statistic: 7.19).

The graph in figure 1 and the statistical test in table 1 cast doubt on the assumption that schools were not able to precisely manipulate their own AYP proficiency targets. The statistical analysis is convincing in part because we were able to develop a reasonably detailed theory of how a school might be able to engage in precise manipulation. Specifically, we argued that if it occurred at all, precise manipulation would likely involve the state exemption rules related to confidence intervals and safe harbors. To follow the logic even further, we conducted more analysis to try to determine whether a single policy was driving the discontinuity in the density function at the cutoff. We reasoned that if a particular exemption rule (i.e. the confidence interval rule) was driving the results of the density test, we would expect to see the discontinuity in the density function only when that rule is applied, but not when other exemption rules (i.e. safe harbor) were used as effective cutoffs.

To examine this hypothesis, we reran the analyses using only the state's confidence interval rule to create site-specific thresholds, and then centered schools' assignment variables on the basis of their new cutoffs. We repeated this exercise two more times by creating site-specific thresholds based solely on the safe harbor rule and then for the confidence

interval around safe harbor rule. Figures 2 and 4 suggest the possibility that schools may be manipulating their proficiency thresholds using the confidence interval rule and possibly the confidence interval around safe harbor rule. The discontinuity in the log difference is large and statistically significant for both (table 1, rows 2 and 4). The figures show a dip in the number of schools immediately before the cutoff followed by a large spike. For the confidence interval cutoff, schools were three times more likely to score at the threshold than below it, and for the confidence interval around the safe harbor target, schools were 2.5 times more likely to just make the cutoff. However, there is no evidence that schools gamed the safe harbor rule (figure 3) given the continuous density function at the cutoff. The discontinuity test indicate—see figure 3—that the log difference here was small (-0.09), and we were not able to reject the Null hypothesis that there was no discontinuity at the cutoff (t-statistic: -0.55) (table 1, row 3).

What should we make of these results? Substantive knowledge about the accountability process for SWD subgroups in Pennsylvania suggest that schools had strong preferences to avoid identification under NCLB, and had complete knowledge of the assignment process. In addition, schools could exercise precise control over their effective proficiency cutoffs through the confidence interval rule. We hypothesized that if schools took advantage of the confidence interval rule to make the AYP cutoff, there should be a dip in the density function below the AYP cutoff, followed by a sharp increase in the number of schools at and above the cutoff. Inspection of AYP data for Pennsylvania schools provided empirical evidence supporting that hypotheses and casting doubt on the null hypothesis that schools could not exercise precise control over their percent proficiency rates. Figure 1 showed a large discontinuity in the densities for schools at the cutoff; figures 2 and 3 suggest that the discontinuity was driven by the confidence interval rules as opposed to the safe harbor rule. Taken together, the results suggest that the RDD estimates of the effects of accountability pressure might underestimate the their true effects, as low performing schools near the cutoff did not actually achieve real gains in student performance. The results also add to the substantive literature on how schools respond to accountability pressures. In this specific instance, it appears that schools were willing to adopt strategies that allowed them to manipulate proficiency thresholds to make AYP cutoffs. There does not appear to be evidence that manipulation occurred around the safe harbor rule, which would require more precise control over the performance of individual students.

6. Example 2: Age Cutoffs for Evaluating the Impacts of State Pre-Kindergarten Programs

A number of studies evaluate early childhood education (ECE) interventions using RDD based on the distribution of school-age cutoffs. The idea is that students with birthdays before the school cutoff are admitted into the ECE program in year one. Students with birthdays after the threshold must wait a year. Treatment effects are assessed by comparing conditional mean outcomes of students with birthdays at the cutoff threshold. Gormley and Phillips (2005) and Gormley et al. (2005) introduced the basic approach in their study of the Tulsa pre-kindergarten (pre-K) program. Wong et al. (2008) used the same basic design to study the effects of five different state pre-K programs. They found that the state pre-K

programs increased print awareness but had smaller and less consistent effects on vocabulary and early math skills.

Subsequent work has raised questions about the internal validity pre-K evaluations using age-related RDDs (Lipsey et al., 2015; Weiland and Yoshikawa, 2013).³ The main worry is that there may be differences between treatment and comparison students near the cutoff that arise from the cross-sectional data collection plan. Another issue is that the birthday distribution may not be smoothly and uniformly distributed across days of the week and months, which could be a problem for some birthday cutoffs under some theories of birthday sorting. In this section, we consider two plausible threats to validity when the assignment process is an age-related cutoff. As in the NCLB example, our goal is to show that many threats to the RDD may be assessed empirically by posing relatively specific theories of manipulation and using those theories to guide the data analysis. For demonstration purposes, our analysis focuses on New Jersey data from a five state pre-K evaluation (Wong et al., 2008).

Pre-K enrollment procedures usually require parents show proof of their child’s birth-date. It seems unlikely that many parents would attempt to manipulate measured birth-dates by fabricating birth certificates to undermine the pre-K entry requirements. However, sorting around the cutoff could arise through other mechanisms as well. Attrition is one concern. In Wong et al. (2008) the outcome measures were collected during the fall of the 2004-05 academic year. That meant that the treatment students were kids who had completed pre-K the prior academic year (2003-04) and were entering public kindergarten when the data were collected. In contrast, the comparison students were kids with birthdays after the school cutoff and were entering public pre-K when the data were collected. Since they measured both groups at the same point in calendar time, Wong et al. did not actually observe the pre-K class of 2003-04 in 2003-04. They observed (possibly) the remnants of the 2003-04 class who were attending public kindergarten in 2004-05. It is possible that some of the treated students who received pre-K in 2003-04 may have transferred to private schools between pre-K and kindergarten. If the movers were from high socioeconomic (SES) families, then one worry is that the surviving treatment group members have systematically lower incomes than their counterparts in the control group.

The ideal design strategy would be to collect information for both treatment and comparison groups across multiple years – students’ baseline information when they enroll in pre-K, and outcome scores when they enter kindergarten. That would make it possible to measure attrition and its correlates directly. However, even with the cross-sectional data collection plan, it is still possible to investigate some of these concerns. A first step is to look for compositional differences between treatment and comparison groups at the cutoff. Under the theory that high SES families systematically transferred to private schools between pre-K and kindergarten, we would expect to see a discontinuity in the proportion of free-reduced-price lunch (FRPL) students at the cutoff. In the null case, the proportion of FRPL students should be continuous across the RD threshold.

3. Lipsey et al. (2015) and Weiland and Yoshikawa (2013) have identified multiple concerns with age-related cutoffs in pre-K settings, including the interpretation of RD treatment effects, comparability of groups at the cutoff, and differences in outcome measurement. In this paper, we focus on empirical methods for assessing comparability of groups at the RD cutoff.

Figure 5 presents evidence on covariate balance at the cutoff for the New Jersey sample of Wong et al. (2008). Children’s birthdays are centered around the school cutoff so that the X-axis measures days away from the cutoff. Pre-K treatment children have birthdays before the cutoff and are represented on the left side of the plot. Comparison children have birthdays after the cutoff and are on the right side of the cutoff. Baseline covariates include: the proportion of students who are African Americans, Hispanic, White, receives free-reduced price lunch, female, and has English as their home language. Figure 5 shows no evidence of discontinuities on baseline covariates at the RD cutoffs. These results provide reassurance that there were no compositional differences between treatment and comparison students at the school cutoff. Still, given that we have only a handful of baseline covariates (race, gender, home language, and free-reduced price lunch status), it is possible that cutoff discontinuities exist on other unobserved characteristics.

As in the NCLB example, we also look for evidence of discontinuities in the estimated density functions at school cutoffs. These plots provide diagnostic clues as to whether there are differences in treatment and comparison groups beyond what was observed by our balance tests on observed covariates. If differential attrition from the treatment group is an issue, density plots should show fewer treatment students exactly above the RD cutoff. Figure 6 provides results from the density test for the New Jersey sample. The dots represent the number of births in birthday bins that radiate out from the cutoff date. The black lines depict smoothed functions of the binned averages based on local linear kernel regression. The McCrary test shows statistical differences in estimated density functions at the cutoff (log difference 0.61, t -statistic = 2.83), where the number of births occurring at or before the school cutoff was 2.3 times larger than just after the cutoff. These results are generally replicated across subgroups. Taken together, the covariate balance analysis and the density analysis are a puzzle. The covariate balance suggests no evidence of differential attrition in the treatment group, but there appears to be a surplus of treatment cases who have birthdays at or above the cutoff.

The extent to which a failed density test is a problem for the RDD depends on the reason for the discontinuity. In the pre-K example, one explanation is that even in the full population, children’s births are not uniformly distributed across months and days of the week. For example, women and families may have preferences for births to occur during the summer months. If monthly birth shares are larger in the summer months, our sample may also include larger shares of treatment children near the cutoff, which include July and September birthdays. At a more granular level, if the school cutoff coincides with a day in which we might expect an abrupt change in the number of births, such as a holiday or even a weekend, there may be a discontinuity in the density of observations around the cutoff. In New Jersey, the birthday cutoff did not occur on a single day, but most pre-K children in the sample (56%) had a school cutoff that fell on Friday, October 1st during their birth year (see table 2). As a result, most of the treatment children near the cutoff would have been born on weekdays, while comparisons just below the cutoff would have been born on the weekend. Figure 8 shows the McCrary density plot for a Tuesday/Thursday cutoff compared to the Friday cutoff. The plot shows support for the weekday hypothesis: there is a large discontinuity in the density of cases when the school cutoff fell on a Friday, but it is relatively smooth for the weekday cutoff.

Sorting behaviors in the birth distribution is an issue for the RDD only in cases when manipulated births have characteristics that are correlated with children’s later achievement outcomes. There are multiple hypotheses for why birthday sorting occurs, but one reason may be that more advantaged women have more flexibility in planning and scheduling their births (Dickert-Conlin and Elder, 2010; Buckles and Hungerman, 2012). Correlations of birth timing and maternal and child characteristics support this claim. Weekday births, and births that occur during non-winter months, are more common among women with higher education levels, married women, and white women (Dickert-Conlin and Elder, 2010; Buckles and Hungerman, 2012). Babies born on weekdays and during non-winter months also have higher birthweights, and are older in gestational weeks (Dickert-Conlin and Elder, 2010; Buckles and Hungerman, 2012). And, Figlio et al. (2014) argue that these family and infant characteristics are related to students’ later SAT scores and academic achievement, suggesting that birthday sorting behaviors may pose validity challenges for RDDs based on age-cutoffs.

Using natality information from the birth year of the pre-K sample, we examined evidence of discontinuities in the underlying New Jersey birth distribution from April 1999 to March 2000, as well as the extent to which day of week and month effects might explain the failed density test in the New Jersey sample. We also considered whether these differences were likely to introduce bias in the RDD treatment effects. While the public use birth certificate records we examined do not provide exact date of births, they do report the month, day of week, and state of each birth. They also include information on the maternal and child characteristics of the births. Overall, birth patterns in New Jersey mirror trends in the general population (Dickert-Conlin and Elder, 2010; Buckles and Hungerman, 2012). In 1999, when our sample of pre-K students were born, there were approximately 17,000 births per weekday, but on weekends, the average dropped to between 11,000 and 12,000 births per day (1999 natality files). In addition, birth rates in New Jersey peaked during non-winter months from May through September. Figures 9 and 10 summarize average maternal and infant characteristics of New Jersey births by month and day of week. The plots show that weekday and summer births were correlated with mothers being white, having college degrees, and being married. Weekday and non-winter infants also had higher birthweights and Apgar scores, and were older in gestational weeks. However, the magnitude of these differences were small – less than 0.10 standard deviations across all covariates, and less than 0.05 standard deviations on most. Combined, these results suggest that if “birth sorting” occurred, the magnitude of the bias is likely to be small but positive with respect to educational achievement.

Finally, using information about the population birth distribution, we constructed a bootstrap test to examine whether the “day of week” and “month” effects did in fact explain the failed McCrary result in the New Jersey sample. Because we did not have access to data on exact dates of birth, we randomly assigned a specific “date” to each New Jersey birth from the set of dates consistent with their observed day of week and month; usually there are four or five possible dates for each birth and we randomly chose a date from this birth specific list. The procedure preserves any “day of week” and “month” effects in the birth distribution, but it would obscure holiday or week effects if they exist. Figure 11 shows the population birth distribution for an October 1, 1999 school cutoff. There is a clear discontinuity in the density function, with approximately 1.2 times more

births before the cutoff than afterwards (log difference = -0.12). We next tested whether the log difference in estimated density functions for the New Jersey pre-K sample on the Friday, October 1st cutoff was larger than what would be expected from sampling error alone, given discontinuities in the underlying birth distribution. For the New Jersey pre-K sample, the log difference at the October 1st cutoff was -0.54. To test whether day of week and month effects explained the difference, we drew 500 random samples of 1,118 students (the size of the New Jersey sample with a Friday, October 1st cutoff) from the population birth distribution. For each sample, we estimated the log difference in density functions at the cutoff to generate a null distribution of effects. We then compared the McCrary test result from our New Jersey sample to the null distribution, and found that only 0.8% of repeated samples produced a McCrary test result as extreme as -0.54. This result suggests that there was more sorting in the New Jersey pre-K data than you would expect given the naturally occurring discontinuities that are present in the overall birth day distribution. It seems that some additional sorting process happened near the October 1st cutoff in the pre-K data.

How should we interpret pre-K results from the New Jersey sample? The covariate balance tests suggest that there were no compositional differences between treatment and comparison groups at the cutoff. However, with only a few demographic covariates (race/ethnicity, free-reduced price lunch status, home language status, and gender), there may be cutoff differences on unobserved attributes that are related to the outcome. Indeed, the density test suggests that some type of sorting did occur at the school cutoff. But why? One hypothesis was that the density test reflected discontinuities in the underlying birth distribution. To examine this hypothesis, we constructed a test based on bootstrap replications from the underlying sample of New Jersey births. That test provided a way to examine whether the pre-K density discontinuity was consistent with underlying discontinuities in the birthday distribution. Test results suggest that the naturally occurring “day of week” and “month” effects did not fully explain the large discontinuity in the density function for the New Jersey pre-K sample. Taken together, researchers should interpret RD treatment effects in New Jersey with caution, at least until an explanation of the failed McCrary test may be ruled out as a threat to the validity of the study.

7. Conclusion

Thistlethwaite and Campbell (1960)’s introduction of the RDD was both comprehensive in its justification and rationalization of the design, as well as prescient in identifying many of the analysis issues that modern researchers would face. Absent in TC’s original presentation of RDD, however, was the concept of manipulation and precise control over assignment scores. As we have argued, these ideas resulted in key diagnostic tools for helping researchers become “Campbellian” in their approach to the RDD. First, develop specific and plausible threats to validity that is based on substantive knowledge of the assignment process. Second, formulate hypotheses of the data pattern that one would expect if the threat was realized. Third, look for empirical evidence of such threats through density tests and covariate balance checks. Finally, if the diagnostic tests fail, consider and test multiple explanations for why there was a discontinuity in the density function. This process helps the researcher

determine if the failed test poses a validity threat to the RDD, and if so, the magnitude and direction of the likely bias.

In cases where the researcher has discovered sorting behaviors that subvert the RD assignment mechanism, the finding may be of interest in its own right, especially when the RD is based on administrative processes that allocate scarce resources. In our No Child Left Behind example, we found evidence that schools manipulated the confidence interval rule for the students with disability subgroup to lower accountability standards. This result has important implications for the design of school accountability policies, as well as for students' legitimate needs for SWD services. The density test may also fail by virtue of the data generating process for creating the assignment variable. For example, an assignment variable may be a test score in which observations heaped on specific values because of how the measure was constructed. In the pre-K RD, we suspected that the failed density test in the New Jersey sample was due to discontinuities in the underlying birth distribution. And in fact, our analysis of the population birth distribution provided evidence of such sorting, which would have likely produced small and positive biases in favor of the treatment. However, our bootstrap test, which took account of discontinuities in the null distribution, suggest that day of week and month effects did not explain fully the discontinuity in the density function in the New Jersey pre-K sample. The conclusion here is that researchers should interpret RD results for the New Jersey sample with caution.

In the pantheon of research designs, TC preferenced the RDD over its more popular cousin, the *ex post facto* design, which was their term for an observational study. Like the RCT, the RDD has a strong theoretical foundation and intuitive appeal. And, as the last 30 years of social science research has demonstrated, the design has many applications for uncovering causal impacts in field settings. But in cases where the treatment is in high demand (or socially undesirable) and the cutoff is well known, the RD is vulnerable to social corruption. Although TC did not address the issue of sorting directly, Campbell's later writings suggest that he was sensitive to the delicate social processes around the use of quantitative indices for decision-making. He writes, "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1976). With thoughtful consideration of study contexts, and careful empirical examination of the data, social corruption of quantitative indicators may be uncovered for both the policy-maker and researcher.

Acknowledgments

The authors wish to thank Dylan Small for comments. All errors are our own.

References

- Ahn, T. and Vigdor, J. (2014). The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina. Technical report, National Bureau of Economic Research, Cambridge, MA. Available from: <http://www.nber.org/papers/w20511.pdf>.
- Buckles, K. S. and Hungerman, D. M. (2012). Season of Birth and Later Outcomes: Old Questions, New Answers. *Review of Economics and Statistics*, 95(3):711–724. Available from: http://dx.doi.org/10.1162/REST_a_00314.
- Campbell, D. T. (1976). Assessing the impact of planned social change. *Occasional Paper Series*, Paper #8.
- Dickert-Conlin, S. and Elder, T. (2010). Suburban legend: School cutoff dates and the timing of births. *Economics of Education Review*, 29:826–841.
- DiNardo, J. and Lee, D. (2010). Program Evaluation and Research Designs. Technical report, National Bureau of Economic Research, Cambridge, MA. Available from: <http://www.nber.org/papers/w16016.pdf>.
- Figlio, D., Guryan, J., Karbownik, K., and Roth, J. (2014). The Effects of Poor Neonatal Health on Children's Cognitive Development. *American Economic Review*, 104(12):3921–3955. Available from: <http://pubs.aeaweb.org/doi/10.1257/aer.104.12.3921>.
- Figlio, D. N. and Getzler, L. S. (2006). Accountability, ability, and disability: Gaming the system. In Gronberg, T. and Jansen, D., editors, *Advances in Microeconomics*. Elsevier, Amsterdam.
- Gill, B., Lockwood, J. R., Martorell, F., Setodji, C. M., and Booker, K. (2008). State and local implementation of the No Child Left Behind Act. Technical report, Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service., Washington, DC.
- Gormley, W. T., Gayer, T., Phillips, D., and Dawson, B. (2005). The Effects of Universal Pre-K on Cognitive Development. *Developmental Psychology*, 41(6):872–884. Available from: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0012-1649.41.6.872>.
- Gormley, W. T. and Phillips, D. (2005). The Effects of Universal Pre-K in Oklahoma: Research Highlights and Policy Implications. *Policy Studies Journal*, 33(1):65–82. Available from: <http://doi.wiley.com/10.1111/j.1541-0072.2005.00092.x>.
- Hahn, J., Todd, P., and Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209.
- Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies*, 79(3):933–959. Available from: <https://ideas.repec.org/a/oup/restud/v79y2012i3p933-959.html>.

- Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2):675–697. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0304407607001121>.
- Lee, D. S. and Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2):281–355. Available from: <http://pubs.aeaweb.org/doi/10.1257/jel.48.2.281>.
- Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., and Hofer, K. G. (2015). The Prekindergarten Age-Cutoff Regression-Discontinuity Design: Methodological Issues and Implications for Application. *Educational Evaluation and Policy Analysis*, 37(3):296–313. Available from: <http://epa.sagepub.com/cgi/doi/10.3102/0162373714547266>.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.
- Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: {An} alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.
- Weiland, C. and Yoshikawa, H. (2013). Impacts of a prekindergarten program on children’s mathematics, language, literacy, executive function, and emotional skills. *Child development*, 84(6):2112–30. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23534487>.
- Wong, V. C., Cook, T. D., Barnett, W. S., and Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, 27(1):122–154. Available from: <http://doi.wiley.com/10.1002/pam.20310>.
- Wong, V. C., Steiner, P. M., and Cook, T. D. (2012). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*.

Figure 1: Smoothed frequency of assignment scores in Pennsylvania schools (SWD subgroup only)

Figure 2: Smoothed frequency of assignment scores with the confidence interval rule as cutoff.

Figure 3: Smoothed frequency of assignment scores with the safe harbor rule as cutoff.

Figure 4: Smoothed frequency of assignment scores with the safe harbor confidence interval rule as cutoff.

Figure 5: Covariate balance at the cut-off in the New Jersey pre-K sample.

Figure 6: Smoothed frequency of birthdates relative to the cut-off in the New Jersey pre-K sample.

Figure 7: Smoothed frequency of birth dates by subgroups in the New Jersey pre-K sample.

Figure 8: Smoothed frequency of birth dates using Tuesday/Thursday vs Friday cutoffs in the New Jersey pre-K sample

Figure 9: Population distribution of births by month (New Jersey, April 1999 to May 2000).

Figure 10: Population distribution of births by days of the week (New Jersey, October 1999).

Figure 11: Manipulation test using the population distribution of New Jersey births as a reference distribution (April 1999 to May 2000).

Table 1: Log discontinuity estimates for NCLB analysis

Discontinuity	Bin size	Bandwidth	$\hat{\theta}$ (SE)	t
Centered cutoff	0.006	0.08	1.51 (.21)	7.19
Confidence Cutoff	0.011	0.14	1.17 (.26)	4.43
Safe Harbor Cutoff	0.005	0.08	-.09 (.15)	-0.55
Safe Harbor-Confidence Interval Cutoff	0.005	0.08	0.95 (.16)	5.94
N		1,030		

Table 2: New Jersey School Cutoff Dates and Sample Sizes

Day of the Week	Cutoff Birth Date	N	Percent
Thursday	9/30/1999	274	13.82
Friday	10/1/1999	1,118	56.41
Friday	10/15/1999	166	8.38
Sunday	10/31/1999	180	9.08
Monday	11/1/1999	103	5.2
Tuesday	11/30/1999	65	3.28
Friday	12/31/1999	76	3.83
Total		1,982	100