Observational Studies 2 (2016) 119-128 Submitted 1960; Published reprinted, 12/16

Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment

Donald L. Thistlewaite and **Donald T. Campbell**

Editor's Note: Donald Thistlewaite (1923-1997) was Professor of Psychology at Vanderbilt University and Donald Campbell (1916-1996) was University Professor at Lehigh University. This article was originally published in *Journal of Educational Psychology*, December 1960, Vol. 51, pp. 309-317. At the time the article was published, Donald Thistlewaite was at the National Merit Scholarship Corporation and Donald Campbell was at Northwestern University. The article is now in the public domain. Comments follow by leading current researchers in regression discontinuity designs: Peter Aronow, Nicole Basta, and Betz Halloran; Matias Cattaneo and Gonzalo Vazquez-Bare; Guido Imbens; Alessandra Mattei and Fabrizia Mealli; Jasjeet Sekhon and Rocío Titiunik; and Vivian Wong and Coady Wing.

1. Introduction

While the term "ex post facto experiment" could refer to any analysis of records which provides a quasi-experimental test of a causal hypothesis, as described by Chapin (1938) and Greenwood (1945), it has come to indicate more specifically the mode of analysis in which two groups – an experimental and a control group – are selected through *matching* to yield a quasi-experimental comparison. In such studies the groups are presumed, as a result of matching, to have been equivalent prior to the exposure of the experimental group to some potentially change inducing event (the "experimental treatment"). If the groups differ on subsequent measures and if there are no plausible rival hypotheses which might account for the differences, it is inferred that the experimental treatment has caused the observed differences.

This paper has three purposes: first, it presents an alternative mode of analysis, called regression-discontinuity analysis, which we believe can be more confidently interpreted than the ex post facto design; second, it compares the results obtained when both modes of analysis are applied to the same data; and, third, it qualifies interpretations of the ex post facto study recently reported in this journal (Thistlethwaite, 1959). Two groups of near-winners in a national scholarship competition were matched on several background variables in the previous study in order to study the motivational effect of public recognition. The results suggested that such recognition tends to increase the favorableness of attitudes toward intellectualism, the number of students planning to seek the MD or PhD degree, the number planning to become college teachers or scientific researchers, and the number who succeed in obtaining scholarships from other scholarship granting agencies. The regression-discontinuity analysis to be presented here confirms the effects upon success in winning

scholarships from other donors but negates the inference of effects upon attitudes and is equivocal regarding career plans.

2. Method

2.1 Subjects and Data[1]

Two groups of near-winners–5,126 students who received Certificates of Merit and 2,848 students who merely received letters of commendation–answered a questionnaire approximately 6 months after the announcement of awards in the second National Merit Scholarship program. The C of M group received greater public recognition: their names were published in a booklet distributed to colleges, universities, and other scholarship granting agencies and they received approximately two and one half times more newspaper coverage than commended students. The decision to award some students the Certificate of Merit, which meant greater public recognition, was made chiefly on the basis of "qualifying scores" on the CEEB Scholarship Qualifying Test (SQT). A second aptitude test, the Scholastic Aptitude Test, was used to confirm the high ability of all finalists, i.e., all students scoring above the SQT qualifying score for the state in which the student attended high school.[2] Two hundred and forty-one students who voluntarily withdrew from the program before the second test or whose scores were not confirmed received neither award while 7,255 students who satisfactorily completed the second test received Certificates of Merit. The latter were subsequently screened by a selection committee and 827 of these students were awarded Merit Scholarships. Since the interest is in estimating the effects of honorary awards, questionnaire responses from Merit Scholars are not included in these analyses. As Table 1 shows, response rate did not vary systematically by test score interval, and there is no reason to believe that differential response bias can account for the effects to be described.

2.2 Regression-Discontinuity Analysis

In situations such as the foregoing, where exposure to an experimental treatment (in this case, increased public recognition) is determined by the subject's standing on a single, measured variable, and where the expected effects of the treatment are of much the same nature as would be produced by increasing magnitudes of that variable, examination of the details of the regression may be used to assess experimental effects. The experimental treatment should provide an additional elevation to the regression of dependent variables on the exposure determiner, providing a steplike discontinuity at the cutting score.

The argument–and the limitations on generality of the result–can be made more specific by considering a "true" experiment for which the regression-discontinuity analysis may be regarded as a substitute. It would be both indefensible and infeasible to conduct an

[1] Details of the sample of students, the experimental treatment, and dependent variables are described in the previous report (Thistlethwaite, 1959), and only the essential features of the data collection will be discussed here

[2] Recognition awards in the 1957 Merit program were distributed so that the number of students recognized in each state was proportional to the number of public high school graduates in each state. Since there were marked state differences in student performance on this test, qualifying scores varied from state to state. All SQT scores represented a composite in which verbal scores were weighted twice as heavily as mathematical scores.

Table 1: Participants in 1957 Merit Program Classified by Aptitude Score Interval

Group	Scholarship qualifying test score interval[b]	Number of Merit Scholars	Number in designated sample[a]	Number of respondents	Percentage of designated sample responding	Percentage of C of M winners in each interval awarded Merit scholarships
Commended students	Below 1		419	322	76.8	
	1		318	256	80.5	
	2		368	281	76.4	
	3		320	258	80.6	
	4		407	338	83.1	
	5		324	259	79.9	
	6		333	267	80.2	
	7		280	213	76.1	
	8		301	248	82.4	
	9		256	201	78.5	
	10		262	205	78.2	
Totals			3,588	2,848	79.4	
Certificate of Merit winners	11	17	476	380	79.8	3.4
	12	22	466	370	79.4	4.5
	13	16	399	319	79.9	3.9
	14	17	371	298	80.3	4.4
	15	19	361	300	83.1	5.0
	16	34	358	289	80.7	8.7
	17	13	319	247	77.4	3.9
	18	18	345	256	74.2	5.0
	19	17	254	211	83.1	6.3
	20	23	301	237	78.7	7.1
	Above 20	631	2,778	2,219	79.9	18.5
Totals		827	6,428	5,126	79.7	11.4

[a] Intervals show the student's SQT score relative to the qualifying score in the student's state, e.g., subjects whose scores equaled the qualifying score are classified in Interval 11, those whose scores were one unit less than the qualifying score are classified in Interval 10, etc.

[b] The designated sample for commended students consisted of a 47% random sample of all commended students.

experiment in which a random group of students along the whole range of abilities would be given the C of M award while a randomly equivalent group received merely the letter of

commendation. However, a group of commended students who narrowly missed receiving the higher award might be given the opportunity of receiving extra recognition. Thus students in Interval 10 in Figure 1 might be randomly assigned to the different treatments of C of M award and no C of M award. The two half-circle points at 10 for Line AA' in Figure 1 illustrate a possible outcome for such a true experiment, the solid half-circle representing the award group, and the hollow half-circle the no award group. Alternatively, a similar true experiment might be carried out among students just above the cutting point (Score 11 in Figure 1). For reasons discussed below, the regression-discontinuity analysis attempts to simulate the latter of these two experiments, by extrapolating from the below-cutting-point line to an "untreated" Point 11 value (an inferred substitute for the no award "control group"). Thus the major evidence of effect must be a distinct discontinuity or difference in intercept at the cutting point. Outcomes such as those shown in Line AA' would, of course, be strictly demonstrated only for aptitude intervals adjacent to the cutting point, and inferences as to effects of the C of M award upon persons of other ability levels would be made in hazard of unexplored interactions of award and ability level. Inferences as to what the regression line would have looked like without the C of M award become more and more suspect the further the no award experience of Points 1 to 10 has to be extrapolated. The extrapolation is best for Point 11 and becomes increasingly implausible for Points 12 through 20.

To better illustrate the argument several hypothetical outcomes are shown in Figure 1. Line AA' indicates a hypothetical regression of the percentage exhibiting Attribute A as a function of score on the decision variable. The steplike discontinuity which begins at the point where the experimental treatment begins to operate would be convincing evidence that the certificate has had an effect upon Attribute A. Similarly, outcomes such as those shown by Lines BB' and CC would indicate genuine treatment effects. Line DD' is a pure case of no effect. Lines EE' and FF' are troublesome: there seems to be a definite change in the regression lines, but the steplike discontinuity at the cutting point is lacking. Consequently the points could merely represent continuous, curvilinear regressions. It seems best not to interpret such ambiguous outcomes as evidence of effects.

In applying this mode of analysis to the present data, the qualifying score in each state was used as a fixed point of reference, and students were classified according to the number of score intervals their SQT score fell above or below the qualifying score in their state. For example, in Figure 2 all students whose scores equaled the qualifying score in their state have been classified in Interval 11, while all those whose scores were one less than the relevant qualifying score have been classified in Interval 10. Data were analyzed only for subjects whose scores placed them within 10 score intervals of the relevant cutting point. Because of nonresponse to particular questionnaire items the *N*s for percentages and means in Figures 2-4 differ slightly from those shown in Column 4 of Table 1.

3. Results

3.1 Graphic Presentation of Results

Figures 2, 3, and 4 present the results for five variables, with least squares linear regression lines fitted to the points. In Figure 2, both regression lines for scholarships received seem to show a marked discontinuity at the cutting point. The persuasive appearance of effect

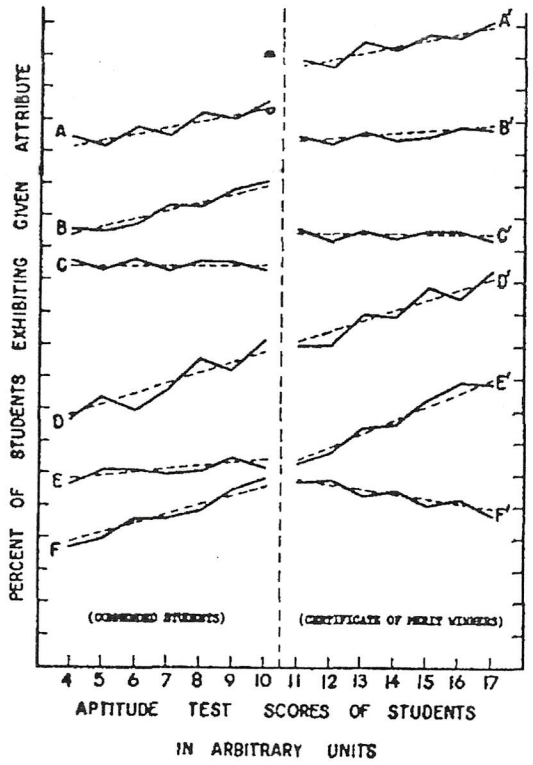

FIG. 1. Hypothetical outcomes of a regression-discontinuity analysis.

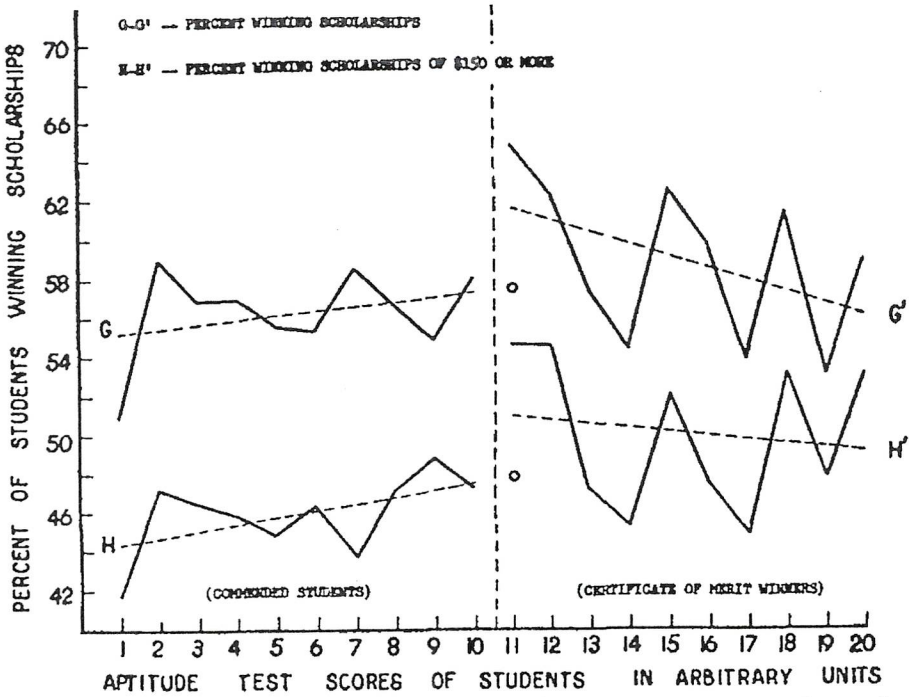

FIG. 2. Regression of success in winning scholarships on exposure determiner.

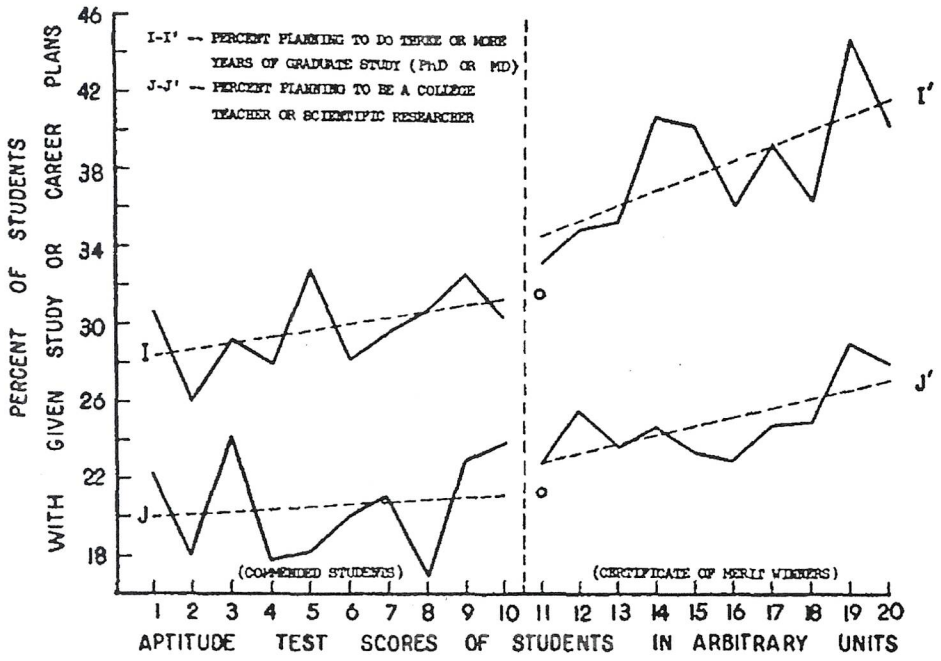

FIG. 3. Regression of study and career plans on exposure determiner.

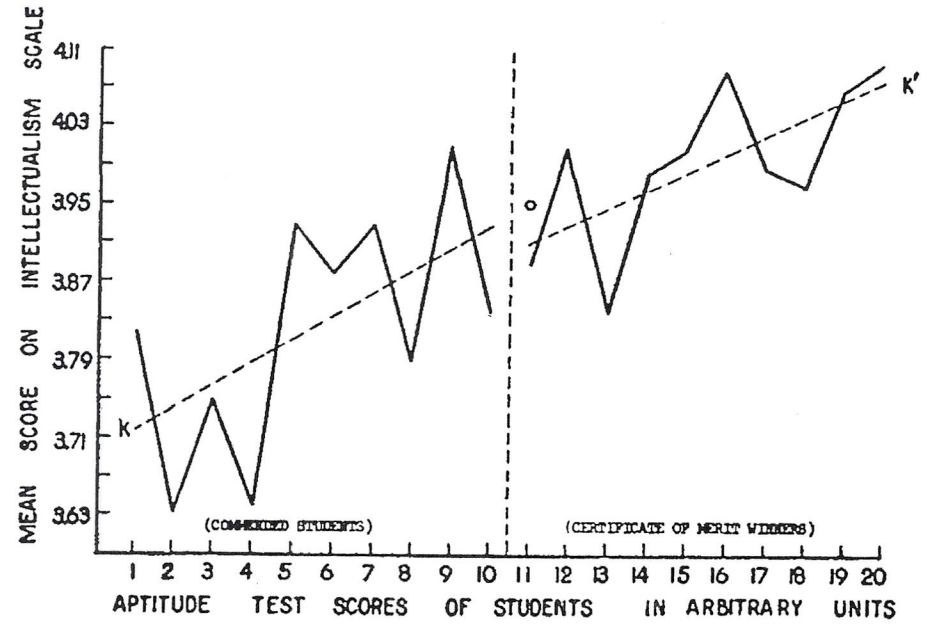

FIG. 4. Regression of attitudes toward intellectualism on exposure determiner.

is, however, weakened by the jaggedness of the regression lines at other points, particularly to the right of the cutting score. In addition, the slopes of the right-hand lines indicate that the effects are specific to students near the cutting score. The downward trend with high scores is presumably a result of eliminating from consideration those receiving Merit Scholarships. Where those of high aptitude test scores are passed over for National Merit Scholarships, it is usually for undistinguished high school grades, which likewise affect the scholarship awards by other agencies as plotted in Figure 2. Table 1 shows that, in general, larger proportions of C of M winners in the highest score intervals were selected for Merit Scholarships.

The two plots in Figure 3 show less discontinuity at the cutting point: there is little or no indication of effect. In II' the difference between observed values at 10 and 11 is small, and while in the hypothesized direction, is exceeded by five other ascending gaps. In JJ' the observed 10-11 jump is actually in the wrong direction. On the other hand, it is confirming of the hypothesis of effect that all of the observed Points 11 through 20 lie above the extrapolated line of best fit for Points 1 to 10, in both II' and JJ'. But this could well be explained by the rival hypothesis of an uninterrupted curvilinear regression from Points 1 to 20. The picture is ambiguous enough to leave us skeptical as to the effects upon the student's study and career plans. The analysis neither confirms nor denies the ex post facto findings.

In Figure 4 no such ambiguity remains. It is inconceivable in view of this evidence that the Certificate of Merit award has increased favorableness of attitudes toward intellectualism, a finding clearly contradicting the ex post facto analysis.

3.2 The Problem of Tests of Significance

In discussing tests of significance in this case, it is probably as important to indicate which tests of significance are ruled out as to indicate those which seem appropriate. Again, reference to the pure cases of Figure 1 will be helpful. A simple t test between Points 10 and 11 is excluded, because it would show significance in an instance like DD' if the overall slope were great enough. That is, such a test ignores the general regression obtained independently of the experimental treatment. Such a test between adjacent points is likewise ruled out on the consideration that even if significant in itself, it is uninterpretable if a part of a very jagged line in which jumps of equal significance occur at numerous other places where not expected. Similarly, a t test of the difference between the means of all points on each side of the cutting point would give significance to cases such as DD' or EE', which would be judged irrelevant. Furthermore, covariance tests applied to the regression lines (e.g., Walker & Lev, 1953, pp. 390-395) are judged inappropriate, because of the differential sample bias for the score intervals arising from the exclusion of Merit Scholars. Even in the ideal case, if the hypothesis of common slope is rejected (as it would be for lines such as EE' and FF') we presumably could not proceed further with a simple linear version of the covariance model.

Mood (1950, pp. 297-298) provides a t test appropriate for testing the significance of the deviation of the first experimental value beyond the cutting point (i.e., the observed Point 11) from a value predicted from a linear fit of the control values (i.e., the encircled point in Figures 2, 3, and 4, extrapolated from Point 1 through 10). As applied here, each plotted

point has been treated as a single observation. On this basis, both of the plots in Figure 2 show a significant effect at Point 11. For GG', $p < .025$; for HH', $p < .01$ (one-tailed tests). Thus the Certificate of Merit seems to have significantly increased chances of obtaining scholarships from other sources. For none of the other figures does this test approach significance. The test in this form fails to make use of the potentially greater stability made available by considering the trend of all of the Values 11 through 20. Potentially the logic of the Mood test could be extended to provide an error term for the difference between two extrapolated points at 10.5, one extrapolated from Points 1 through 10, the other from Points 11 through 20. In many applications of the regression discontinuity analysis, this would be the most appropriate and most powerful test. In our present instance, we have judged it inappropriate because of the differential sampling bias felt to exist in the range of Points 11-20, as explained above.

4. Discussion

A critic may easily question the results of an ex post facto experiment by supposing that one or more relevant matching variables has been inadequately controlled or entirely overlooked. In contrast the regression discontinuity analysis does not rely upon matching to equate experimental and control groups, hence it avoids the difficulties of (a) differential regression toward-the-mean effects, and (b) incomplete matching due to failure to identify and include all relevant antecedent characteristics in the matching process.

Edwards (1954, pp. 279-282) has shown how pseudo effects may be produced in ex post facto designs through differential regression effects. Suppose, for example, we were to match, with respect to aptitude test scores, a group exposed to recognition and a group not exposed to recognition. Since exposure to recognition tends to be positively correlated with aptitude test score we expect that the matched experimental subjects will have low aptitude scores relative to other exposed subjects, while the matched control subjects will have high aptitude scores relative to other unexposed subjects. To the extent that there are errors of measurement on the aptitude variable, however, our experimental group is apt to contain subjects whose aptitude scores are too low through error, while our control group is apt to contain subjects whose aptitude scores are too high through error. Simply on the basis of regression effects, then, we can predict that the matched experimental group will excel the matched control group on a subsequent administration of the aptitude test and on any other variable positively correlated with aptitude. Following Thorndike (1942, pp. 100-101), who discussed a similar problem, one might attempt to match individuals on the basis of predicted true score on the background traiti.e., score predicted by the regression equation between original test and a retest at the time of the experimental comparison. However, the predicted true score for each individual must be determined from the regression equation for his own population, and for groups when the special treatment is not applied. Unfortunately such matching is usually impossible in situations where we wish to use the ex post facto design, since we typically cannot obtain pretest and posttest measures on control variables for "experimental" groups from which the special treatment has been withheld. Indeed if we had the power to withhold the treatment from some subjects we would usually be able to test our causal hypotheses by an experiment with true randomization. In short, the

suggested procedure for controlling regression effects in ex post facto studies presupposes knowledge which we typically cannot obtain.

In the present analysis exposed and unexposed groups are subdivided according to their closeness to receiving a treatment other than the one they have received. Background traits correlated with the probability of exposure to recognition (e.g., rank in high school graduating class, scholastic aptitude, etc.) presumably vary systematically with the score intervals which represent the student's nearness to the cutting point. All of these traits contribute to the observed slopes of the regression lines plotted in Figures 2-A. Since there is no reason to believe that the composite effect of all relevant background traits fluctuates markedly at the cutting point, regression discontinuities emerging at the 10-11 gap must be attributable to the special experimental treatment-the only factor which assumes an abrupt change in value in this region. Thus the new analysis seems to provide a persuasive test of the presence or absence of experimental effects.[3]

The value of the regression-discontinuity analysis illustrated here is that it provides a more stringent test of causal hypotheses than is provided by the ex post facto design. Admittedly the class of situations to which it is applicable is limited. This class consists of those situations in which the regression of dependent variables on a single determiner of exposure to an experimental treatment can be plotted. Whenever the determiners of exposure are multiple or unknown this mode of analysis is not feasible. Of the five variables described in Figures 2-4 the regression-discontinuity analysis indicated significant effects only for those shown in Figure 2. The ex post facto experiment, on the other hand, indicated significant effects for all variables except HH' (success in winning a freshman scholarship of $50 or more). For six other variables, not reported here, neither analysis indicated a significant effect.[4] Considering the regression-discontinuity analysis to be the more definitive, it appears that the ex post facto experiment underestimated effects for one variable and wrongly indicated effects for three variables.

We conclude that increased public recognition tends to increase the student's chances of winning scholarships. There is no clear-cut evidence in the present analysis that such recognition affects the student's career plans, although an effect upon plans to seekgraduate or professional degrees is not ruled out. In this regard, Thistlethwaite (1961) has reported that when near-winners in a subsequent National Merit program were asked, "How did winning a C of M help you?" approximately two out of every five reported that it "increased my desire for advanced training (MA, PhD, MD, etc.)." In short, while other evidence indicates that the hypothesis of effect upon study plans may be correct, the present analysis does not provide confirmation.

[3]Background traits uncorrelated with the probability of exposure to recognition will, of course, not vary systematically with score intervals, but these traits are irrelevant. Even if partialed out they would not affect the correlation between the dependent variable and degree of exposure to recognition.

[4]No significant differences were found with respect to the percentages enrolling in college immediately, well satisfied with their choice of college, believing their college offers the best training in their field of study, going to college more than 250 miles from home, applying for two or more scholarships, or receiving encouragement from their high school teachers and guidance counselors to go to college.

5. Summary

The present report presents and illustrates a method of testing causal hypotheses, called regression-discontinuity analysis, in situations where the investigator is unable to randomly assign subjects to experimental and control groups. It compares the results obtained by the new mode of analysis with those obtained when an ex post facto design was applied to the same data. The new analysis suggested that public recognition for achievement on college aptitude tests tends to increase the likelihood that the recipient will receive a scholarship but did not support the inference that recognition affects the student's attitudes and career plans.

References

Chapin, F.S. (1936) Design for social experiments. *American Sociological Review*, 3: 786-800.

Edwards, A.L. (1954). Experiments: Their planning and execution. In G. Lindzey (Ed), *Handbook of social psychology.* Vol. 1. Cambridge, Mass : Addison-Wesley.

Greenwood, E. (1945). *Experimental sociology: A study in method.* New York: King's Crown.

Mood, A.M. (1950). *Introduction to the theory of statistics.* New York: McGraw-Hill.

Thistlewaite, D.L. (1959). Effects of social recognition upon the educational motivation of talented youth. *Journal of Educational Psychology*, 50: 111-116.

Thistlewaite, D.L. (1961). The recognition of excellence. *College and University*, 36: 282-295.

Thorndike, R.L. (1942). Regression fallacies in the matched group experiment. *Psychometrika*, 7: 85-102.

Walker, H.M. and Lev, J. (1953). *Statistical Inference.* New York: Holt.

Observational Studies 2 (2016) 129-133 Submitted 10/16; Published 12/16

The Regression Discontinuity Design Under Interference: A Local Randomization-based Approach

Peter M. Aronow **peter.aronow@yale.edu**
Department of Political Science, Department of Biostatistics and Institution for Social and Policy Studies
Yale University, New Haven, CT, U.S.A.

Nicole E. Basta **nebasta@umn.edu**
Division of Epidemiology and Community Health, School of Public Health
University of Minnesota, Minneapolis, MN, U.S.A

M. Elizabeth Halloran **betz@fhcrc.org**
Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center and
Department of Biostatistics, University of Washington
Seattle, WA, U.S.A.

1. Introduction

In recent years, the regression discontinuity (RD) design originally pioneered in educational psychology (Thistlethwaite and Campbell, 1960) has been rediscovered and has been the focus of much methodological development (see, e.g., Lee, 2008; Imbens and Lemieux, 2008). One particularly promising avenue of research, typified by Cattaneo, Frandsen, and Titiunik (2015), considers a local randomization-based model for the RD design. In the local randomization-based approach to the RD design, it is hypothesized that, within some finite *window* of an administrative threshold (e.g., a test score or age cutoff) that determines treatment assignment, subjects are "as-if" randomly assigned to treatment and control.

Despite recent methodological advances for the RD design, comparatively little attention has been given to the cases where there may be interference (Cox, 1958) between subjects (i.e., some subjects' treatment status may causally affect other subjects' outcomes). One notable exception is Cattaneo, Titiunik, and Vazquez-Bare (2016), which provides a software implementation of Rosenbaum (2007)'s interference-robust confidence intervals for Fisher (1935)-type causal inference with the RD design under a local randomization assumption. In this short note, we investigate the properties of the RD design for Neyman (1923)-type causal inference under a simple local randomization-based model when we allow for interference of arbitrary and unknown structure. We show that under a local randomization assumption, the difference-in-means estimator as applied to subjects in the window near the cutoff is unbiased for a regime-specific causal effect. This causal effect is equivalent to Hudgens and Halloran (2008)'s average direct effect for the subpopulation of subjects within the window near the threshold. For topics of study where interference is likely widespread

(e.g., evaluating effects of vaccines), our result may help to provide a formal basis for the use of the RD design.

2. Results

We first define our causal model and setting. Suppose we have a finite population U of N subjects indexed by $i = 1, ..., N$. Define a treatment assignment vector, $\mathbf{z} = (z_1, ..., z_N)'$, where $z_i \in \{0, 1\}$ specifies which treatment value (0 = control, 1 = treatment) that subject i receives. Suppose that associated with each subject i are 2^N fixed potential outcomes, $y_i(\mathbf{z})$, defined over all 2^N vectors $\mathbf{z}$ such that $z_j \in \{0, 1\}, \forall j \in \{1, 2, .., N\}$. For example, $y_2((0, 1, 0, ..., 0))$ would represent subject 2's potential outcome if she were treated, but no other subjects were treated. Note that this makes no assumptions about interference (or its absence): for each subject i, her outcome y_i is allowed to depend on her own or any combination of the other subjects' treatment statuses.

To proceed, we will make assumptions about the manner in which $\mathbf{z}$ is assigned and how some potential outcomes are revealed. In particular, we will assume that $\mathbf{Z}$ is a binary random vector of length N. The observed data then consist of a single realization from $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$, where it is assumed that the outcome vector $\mathbf{Y} = (y_1(\mathbf{Z}), y_2(\mathbf{Z}), ..., y_N(\mathbf{Z}))$. ($\mathbf{X}$ is discussed in the following paragraph; note that potential outcomes $y_i(\mathbf{z})$ are assumed not to depend on $\mathbf{X}$.) The key idea behind the local randomization-based approach is to assume that, for a range of subjects with values of a "running variable" (e.g., age in days) that are close to the administrative threshold (e.g., the minimum age for which treatment is administered), treatment is administered in an as-if random manner. We formalize this notion as follows.

Assume that associated with each subject i is a random variable X_i denoting the difference between subject i's running variable and the threshold. We assume that whether or not a subject is treated depends solely on whether or not she is above the threshold: let $Z_i = 1$ (the subject is treated) if $X_i \geq 0$ (the subject is above the threshold), else let $Z_i = 0$ (the subject is in the control condition). In the event that we have noncompliance, let Z_i instead denote treatment eligibility and all effects should be interpreted as intention-to-treat effects. The joint distribution of $\mathbf{X} = (X_1, X_2, ..., X_N)$ fully determines the joint distribution of $\mathbf{Z}$, hence knowledge of $\mathbf{X}$ implies knowledge of $\mathbf{Z}$. Here we introduce our key modeling assumption, so as to be consistent with the local randomization approach. Let b be the bandwidth, or the maximum distance (e.g., number of days) from the threshold at which we would still consider a subject to be close to the threshold. Denote the window $W_0 = [-b, b]$, and define the conditional distribution function $F_{\mathbf{x},W_0}$ as the empirical distribution function of X_i for all subjects such that $X_i \in W_0$. To simplify our definition of local randomization, without loss of generality, assume that $\Pr(X_i = X_j) = 0, \forall i, j : X_i, X_j \in W_0$, so that no two observed age values within the window are exactly identical. Our local randomization assumption requires that $F_{\mathbf{x},W_0}(t) = F_0(t), \forall \mathbf{x} \in \text{Supp}(\mathbf{X})$ and $\forall t$; $\Pr(X_i \in W_0) \in \{0, 1\}, \forall i \in \{1, ..., N\}$, $\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{N_0!}, \forall \mathbf{x} \in \text{Supp}(\mathbf{X})$, where $N_0 = \sum_{i=1}^{N} \mathrm{I}(X_i \in W_0)$: implying that all permutations of the running variable values (and therefore treatment values) within the window W_0 are equiprobable. Further assume that $0 < \sum_{i:X_i \in W_0} Z_i < N_0$, so that at least one subject within the window is treated, and at least one subject within the window is in the control group.

We now define our causal estimands. Denote the individual average potential outcome under treatment z for all subjects $i : X_i \in W_0$,

$$\overline{Y}_i(z) = \frac{\sum_{\mathbf{z} \in \text{Supp}\,(\mathbf{Z})} y_i(\mathbf{z}) \text{I}(\mathbf{z}_i = z)}{\sum_{\mathbf{z} \in \text{Supp}\,(\mathbf{Z})} \text{I}(\mathbf{z}_i = z)}.$$

Note that this estimand is regime specific: it generally depends on the joint distribution of $\mathbf{Z}$. Depending on the treatment allocation scheme, then the individual average potential outcome would change. (E.g., when 90% of subjects in the population are treated, then both $\overline{Y}_i(0)$ and $\overline{Y}_i(1)$ might be different from when 10% of subjects in the population are treated.) While the treatment allocation scheme for subjects outside of W_0 is left completely unspecified, the allocation scheme for these subjects nevertheless contributes to the definition of each $\overline{Y}_i(z)$. Individual average potential outcomes marginalize over the treatment assignments for all units, not simply those within the window near the threshold.

Then our primary target is the average direct effect (Hudgens and Halloran, 2008) of treatment for subjects who are close to the threshold:

$$\tau = N_0^{-1} \sum_{i:X_i \in W_0} \overline{Y}_i(1) - \overline{Y}_i(0).$$

Or, put simply, our target is the difference between the average outcome we would expect to see in treated subjects (across all subjects within the window and across all randomizations) and the average outcome we would expect to see in control subjects (across all subjects within the window and across all randomizations). When there is interference between subjects, this estimand is conditional not only on the subjects being close to the threshold in the running variable, but also on the exact way in which the treatment is assigned to all subjects in the finite population U.

We now show that τ is estimable and that the difference-in-means estimator as applied to all subjects $i : X_i \in W_0$ is unbiased for the average direct effect τ among these subjects. Formally, the difference-in-means estimator for subjects local to the threshold,

$$\hat{\tau} = \frac{\sum_{i:X_i \in W_0} Y_i Z_i}{\sum_{i:X_i \in W_0} Z_i} - \frac{\sum_{i:X_i \in W_0} Y_i (1 - Z_i)}{\sum_{i:X_i \in W_0} (1 - Z_i)}.$$

Then, under the assumption of local randomization, the difference-in-means estimator as applied to subjects local to the threshold is unbiased for the average direct effect for subjects near the threshold:

$$\text{E}\left[\hat{\tau}\right] = \frac{1}{N_0} \left[\sum_{i:X_i \in W_0} \overline{Y}_i(1) - \sum_{i:X_i \in W_0} \overline{Y}_i(0) \right],$$

where the result trivially follows from linearity of expectations.

Inference on the average direct effect for subjects near the threshold requires considerably more structure. As a sufficient condition for root-n consistency, asymptotic normality, and the existence of conservative variance estimators, asymptotics based on a growing number of strata of bounded size suffice, given (i) partial interference (i.e., subjects do not

interfere with subjects outside of their stratum, see Sobel 2006) and (ii) suitable regularity conditions on the values of potential outcomes, $y_i(\mathbf{z})$, and the within-stratum running variable distributions, $F_0(t)$. An alternative sufficient condition, without requiring a large number of independent strata, is given by stratified interference and suitable regularity conditions. Liu and Hudgens (2014) and Aronow and Samii (2016) provide details. One further alternative, as suggested by Cattaneo, Frandsen, and Titiunik (2015) and Cattaneo, Titiunik, and Vazquez-Bare (2016), is to impose more structure on causal effects and use a variant of Fisher (1935)'s exact test, which may be preferable particularly when N_0 is small.

3. Discussion

Our results have illustrated that the RD design can yield credible causal inferences in the context of studies with interference, though — as is usually the case under interference — the estimated causal effect may have a nuanced interpretation. We note here three possible avenues for future methodological work in the context of interference, including (i) exploring results analogous to ours in the setting of the standard econometric limit-based RD design (Imbens and Lemieux, 2008); (ii) derivation of the properties of "fuzzy" instrumental variables-type estimators that account for imperfect compliance (Angrist and Lavy, 1999; Angrist and Pischke, 2009); (iii) exploiting natural variation in local treatment saturation to estimate indirect, total and overall effects as in Hudgens and Halloran (2008).

Acknowledgments

The authors acknowledge support from NIAID grants R37-AI032042 and R01-AI085073. The authors thank Jonathon Baron, William-Jack Dalessandro, Molly Offer-Westort, and Rocio Titiunik for helpful discussions.

References

Angrist, J.D. and V. Lavy (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*. 114(2): 533 – 575.

Angrist, J.D., and J.S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Aronow, P.M. and C. Samii (2016). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, in press.

Cattaneo, M.D., B.R. Frandsen, and R. Titiunik (2015). Randomization inference in the regression discontinuity design: an application to party advantages in the U.S. Senate." *Journal of Causal Inference*. 3(1): 1 –24.

Cattaneo, M.D., R. Titiunik, and G. Vasquez-Bare (2016). Inference in regression discontinuity designs under local randomization. *The Stata Journal*. 16(2): 331 – 367.

Cox, D.R. (1958). *Planning of Experiments*. Oxford, UK: Wiley.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Hudgens, M.G. and M.E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association*. 103(482): 832 – 842.

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

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

Liu, L. and M.G. Hudgens (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*. 109(505): 288 – 301.

Neyman, J.S. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. Translated and reprinted by
erzy Splawa, D. M. Dabrowska, and T. P. Speed [1923.] 1990. On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science* **5**: 465–480.

Rosenbaum, P.R (2007). Interference between units in randomized experiments. *Journal of the American American Statistical Association*. 102(477): 191 – 200.

Sobel, M.E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association* 101 (476): 1398–1408.

Thistlethwaite, D. and D. Campbell. (1960). Regression-discontinuity analysis: an alternative to the ex post facto experiment." *Journal of Educational Psychology*. 51: 309 – 317.

Observational Studies 2 (2016) 134-146 Submitted 8/16; Published 12/16

The Choice of Neighborhood in Regression Discontinuity Designs

Matias D. Cattaneo **cattaneo@umich.edu**
Department of Economics and Department of Statistics
University of Michigan
Ann Arbor, MI 48104, US

Gonzalo Vazquez-Bare **gvazquez@umich.edu**
Department of Economics
University of Michigan
Ann Arbor, MI 48104, US

This version: January 22, 2017

The seminal paper of Thistlethwaite and Campbell (1960) is one of the greatest breakthroughs in program evaluation and causal inference for observational studies. The originally coined *Regression-Discontinuity Analysis*, and nowadays widely known as the Regression Discontinuity (RD) design, is likely the most credible and internally valid quantitative approach for the analysis and interpretation of non-experimental data. Early reviews and perspectives on RD designs include Cook (2008), Imbens and Lemieux (2008) and Lee and Lemieux (2010); see also Cattaneo and Escanciano (2017) for a contemporaneous edited volume with more recent overviews, discussions, and references.

The key design feature in RD is that units have an observable running variable, score or index, and are assigned to treatment whenever this variable exceeds a known cutoff. Empirical work in RD designs seeks to compare the response of units just below the cutoff (control group) to the response of units just above (treatment group) to learn about the treatment effects of interest. It is by now generally recognized that the most important task in practice is to select the appropriate neighborhood near the cutoff, that is, to correctly determine which observations near the cutoff will be used. Localizing near the cutoff is crucial because empirical findings can be quite sensitive to which observations are included in the analysis. Several neighborhood selection methods have been developed in the literature depending on the goal (e.g., estimation, inference, falsification, graphical presentation), the underlying assumptions invoked (e.g., parametric specification, continuity/nonparametric specification, local randomization), the parameter of interest (e.g., sharp, fuzzy, kink), and even the specific design (e.g., single-cutoff, multi-cutoff, geographic).

We offer a comprehensive discussion of both deprecated and modern neighborhood selection approaches available in the literature, following their historical as well as methodological evolution over the last decades. We focus on the prototypical case of a continuously distributed running variable for the most part, though we also discuss the discrete-valued case towards the end of the discussion. The bulk of the presentation focuses on neighborhood selection for estimation and inference, outlining different methods and approaches according to, roughly speaking, the size of a typical selected neighborhood in each case, going from the largest to smallest neighborhood. Figure 1 provides a heuristic summary, which we

Figure 1: Choice of neighborhood (single-cutoff sharp RD)

discuss in detail throughout this article. This ordering among neighborhood selectors is not strict, but it does reflect typical empirical results and may hold under reasonable assumptions and conditions. Furthermore, this ordering follows roughly the historical evolution in the empirical and methodological RD literatures. To complement the discussion, we also reflect briefly on neighborhood selection for several falsification and validation approaches that have recently been proposed in the RD literature.

Our main methodological discussion and recommendations apply not only to the most standard single-cutoff sharp RD design but also more generally to many other RD settings such as fuzzy RD designs (e.g., Hahn et al., 2001), kink RD designs (e.g., Card et al., 2015, 2017), geographic RD designs (e.g., Keele and Titiunik, 2015; Keele et al., 2017), multi-cutoff RD designs (e.g., Cattaneo et al., 2016b), derivative estimation and stability testing (e.g., Dong and Lewbel, 2015; Cerulli et al., 2017), distributional treatment effects (e.g., Shen and Zhang, 2016), and density discontinuity designs (e.g., Jales and Yu, 2017). Adapting the main discussion to these other RD settings is not difficult because our main methodological points are conceptual, and hence not directly tied to any specific RD setup (i.e., only the underlying technicalities or specific features of the problem considered would change, not the general message).

The last section summarizes the implications of our methodological points in the form of concrete recommendations for practice. This section builds on the most recent, and still rapidly expanding, methodological literature on RD designs. Our recommendations are given in general terms so they can be followed in most, if not all, empirical settings employing any regression discontinuity design.

Choosing a Neighborhood

No matter the approach taken (parametric, nonparametric, local randomization) or specific goal (estimation, inference, falsification, graphical presentation) when selecting a neighborhood around the RD cutoff, researchers must impose assumptions, explicitly or implicitly, which they deem reasonable and applicable for the empirical problem at hand. Therefore, it is rarely the case that a method strictly dominates everything else: at the core of the underlying reasoning often lays a trade off between efficiency and robustness, where some methods will be more "efficient" under the assumptions imposed, but more sensitive to violations of these assumptions, while other methods will be more "robust" to such violations but usually at the cost of some loss in precision.

We do rank approaches because we take a stand on the efficiency-robustness trade off: since empirical researchers never know the features of the underlying data generating process, and pre-testing for such features (when possible) can lead to other methodological and practical problems in terms of estimation and inference, we favor procedures that are valid under weaker assumptions, that is, we prefer more *robust* methods. From this robustness perspective, a clear ranking among most neighborhood selectors emerges naturally, as we discuss precisely in this section.

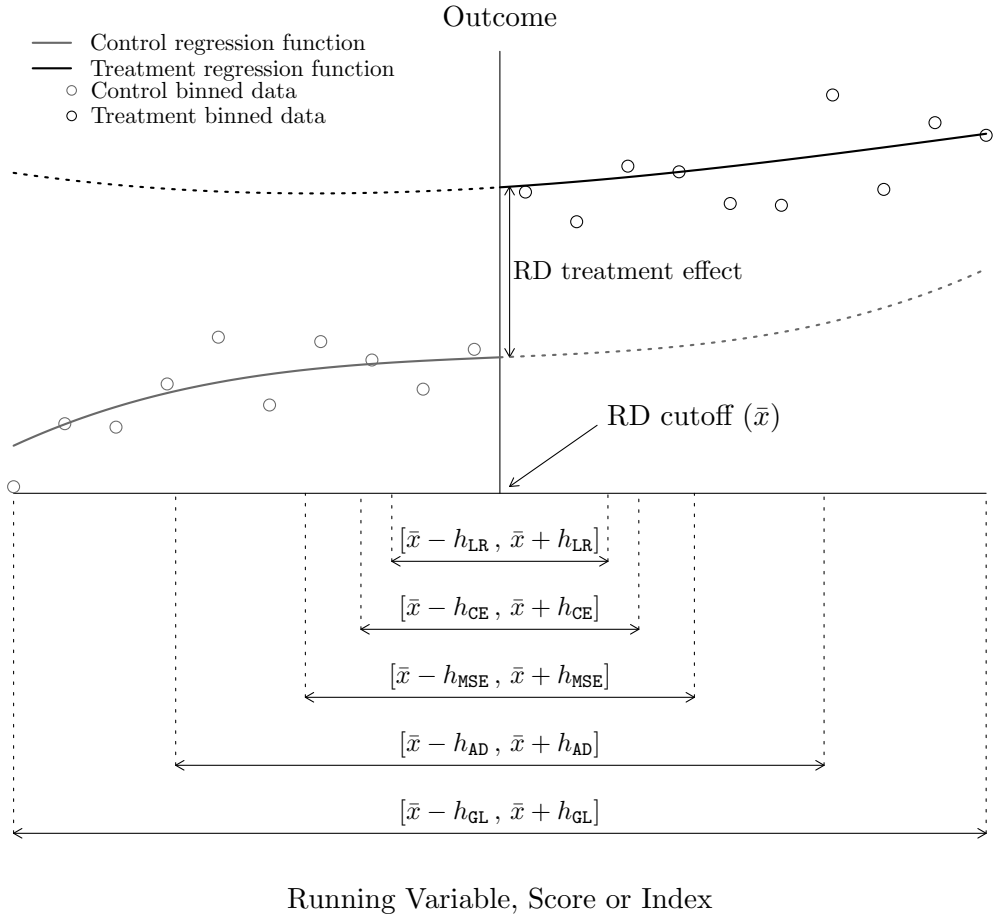

Figure 1: Choice of neighborhood (single-cutoff sharp RD)

Ad-Hoc Neighborhood

We classify as ad-hoc neighborhood selection those approaches that do not employ the data at all to select the neighborhood or, at least, not in a systematic and objective way. These methods were quite popular in the early stages of the RD design developments, but are nowadays widely viewed as inappropriate for the analysis and interpretation of RD designs. We discuss them here not only because they were the first used, but also because they give a natural introduction to the modern approaches outlined further below.

The very first (ad-hoc) method for selecting a neighborhood around the RD cutoff was to employ the full support of the data together with a linear regression model for estimation and inference, which traces back all the way to Thistlethwaite and Campbell (1960). Later, once the crucial role that global extrapolation plays in this approach was fully appreciated, practitioners moved towards either (i) selecting a "smaller" neighborhood in an arbitrary way (and still use linear regression), or (ii) employing higher-order polynomial regression (and still use the full support of the data). These two approaches were popular for some time in early empirical work employing RD designs. Figure 1 offers a graphical schematic of these methods: $h_{\texttt{GL}}$ stands for the "global" or full support approach, where usually a higher-order polynomial is used, and $h_{\texttt{AD}}$ denotes the ad-hoc "local" neighborhood, where the researcher chooses the bandwidth in arbitrary manner. This smaller ad-hoc, parametric linear regression neighborhood is depicted as "large" relative to other modern methods discussed below because in our experience most empirical applications and/or real datasets we have reanalyzed employing the latter methods typically exhibited this pattern. In other words, ad-hoc neighborhoods were usually chosen to be large relative to what automatic, data-driven methods would have selected instead.

Obvious concerns with methods that select a neighborhood around the RD cutoff in an ad-hoc way are: (i) lack of objectivity, (ii) lack of comparability, and (iii) lack of control over the researcher's discretion. In contrast, all of the data-driven procedures that we discuss below avoid these issues, and hence they provide at least a useful benchmark for empirical work exploiting regression discontinuity designs.

Another important, but more subtle, worry related to ad-hoc neighborhood selection methods relates to the underlying assumptions imposed when conducting estimation and inference, which many times are not even explicitly acknowledged by practitioners. To be specific, underlying any of the ad-hoc methods commonly encountered in empirical work there is a crucial assumption: the regression function is correctly specified or, at least, any misspecification error is small enough to be ignored. This parametric approach to RD designs gives practitioners justification to employ standard least squares results when conducting estimation and inference. While such parametric approach is, of course, correct when the regression functions are correctly specified, in general there is no reason for the *unknown* conditional expectations to have the exact (or close enough) parametric form postulated, and hence misspecification errors can be a serious concern. Furthermore, it is now well recognized that employing higher-order polynomial approximations over a large support is highly detrimental, when the goal is to learn something about a boundary point as in RD designs, because such an approach leads to counterintuitive weighting of observations (Gelman and Imbens, 2014) and erratic behavior of the estimator near the boundary (usually known as the Runge's phenomenon, see Calonico et al., 2015, for more discussion).

Finally, some empirical researchers have used ad-hoc neighborhood selectors based on data-driven procedures from the nonparametric literature, such as those related to bandwidth selection for kernel-based density estimation (Wand and Jones, 1995) or local polynomial estimation at an interior point (Fan and Gijbels, 1996). While these approaches are data-driven, they are also ad-hoc in sense that they are not tailored to RD designs, and hence they can lead to invalid (or at least suboptimal) estimation and inference procedures. These approaches are not very popular in modern empirical work employing RD designs, nor are they recommended or theoretically justified, and therefore we do not discuss them further.

The concerns and criticisms outlined above have led modern researchers to employ fully data-driven, objective neighborhood selectors to conduct estimation and inference in RD designs. Ad-hoc methods are nowadays deprecated and dismissed among most well trained practitioners and methodologists. If used, they are typically presented as supplementary evidence after reporting results based on the data-driven methods discussed next, which enjoy demonstrably optimality and/or robustness properties.

Local Polynomial Neighborhood: MSE-Optimal Point Estimation

In this and related approaches, the neighborhood takes the form $[\bar{x}-h, \bar{x}+h]$ and hence is determined by a choice of bandwidth h. Imbens and Kalyanaraman (2012, IK hereafter) were the first to propose an objective neighborhood selector specifically tailored for RD designs. They developed a Mean Squared Error (MSE) optimal bandwidth choice for the local-linear regression point estimator in sharp and fuzzy RD designs. This result was later extended to (i) general local polynomial point estimators, (ii) kink RD designs, (iii) clustered data, (iv) inclusion of pre-intervention covariates, and (v) different bandwidth choices on the left and on the right of the cutoff, in a sequence of more recent papers (Calonico et al., 2014; Bartalotti and Brummet, 2017; Calonico et al., 2016c).

The MSE-optimal bandwidth takes the form $h_{\texttt{MSE}} = C_{\texttt{MSE}} \cdot n^{-1/(2p+3)}$, where n denotes the total sample size available, p denotes the polynomial order used for estimation ($p = 1$ for linear regression), and the constant $C_{\texttt{MSE}}$ involves several known and unknown quantities that depend on objects such as the kernel function, p, the parameter of interest, the asymptotic bias and variance of the estimator, the evaluation point (in multi-cutoff or geographic RD designs), and even whether additional pre-intervention covariates were included in the estimation. This approach is also depicted in Figure 1.

Given a sample size n, the infeasible MSE-optimal neighborhood $[\bar{x}-h_{\texttt{MSE}}, \bar{x}+h_{\texttt{MSE}}]$ will be larger as the value of the unknown constant $C_{\texttt{MSE}}$ increases. This constant, in turn, will become larger whenever the variability of the estimator and/or model increases near the cutoff (e.g., p is larger, the conditional variance of the outcome is larger or the density of observations near the cutoff is smaller) and whenever the parametric approximation improves near the cutoff (i.e., less misspecification bias). In practice, $h_{\texttt{MSE}}$ is constructed by first forming a preliminary estimator $\hat{C}_{\texttt{MSE}}$ of the unknown constant $C_{\texttt{MSE}}$, leading to the estimated bandwidth $\hat{h}_{\texttt{MSE}} = \hat{C}_{\texttt{MSE}} \cdot n^{-1/(2p+3)}$, and therefore the selected neighborhood around the RD cutoff $\bar{x}$ takes the form $[\bar{x}-\hat{h}_{\texttt{MSE}}, \bar{x}+\hat{h}_{\texttt{MSE}}]$. IK proposed a first-generation plug-in rule leading to a bandwidth selector $\hat{h}_{\texttt{MSE}}$, based on a simple reference model and (possibly inconsistent) plug-in estimators. An improved, second-generation bandwidth selector was

later developed by Calonico et al. (2014, 2016c), which enjoys demonstrably superior finite and large sample properties relative to the original IK's bandwidth selector. See, e.g., Wand and Jones (1995) for a discussion of first- and second-generation bandwidth selectors, and their statistical properties.

In this MSE-optimal point estimation approach, only observations with their running variable laying within the selected neighborhood $[\bar{x} - \hat{h}_{\texttt{MSE}}, \bar{x} + \hat{h}_{\texttt{MSE}}]$ are used for estimation of the RD treatment effect. This estimator is fully data-driven, objective and optimal in a mean squared error sense, which makes it highly desirable for empirical work, at least as a benchmark estimate. Employing second-generation plug-in bandwidth selectors lead to superior performance of the MSE-optimal RD treatment effect estimator in finite and large samples.

At the same time, the MSE-optimal point estimator cannot be used directly for inference, that is, for constructing confidence intervals, conducting hypothesis tests or assessing statistical significance. At the core of the argument lays a fundamental logical inconsistency: the neighborhood $[\bar{x} - \hat{h}_{\texttt{MSE}}, \bar{x} + \hat{h}_{\texttt{MSE}}]$ is selected for MSE-optimal point estimation and hence balances bias-squared and variance in a way that makes, by construction, inference invalid when the same observations and RD estimator are used. There is no way out of this logical inconsistency: if one assumes that the misspecification bias is not present (i.e., $\mathsf{bias} = 0$), then $h_{\texttt{MSE}}$ is necessarily not well defined because $C_{\texttt{MSE}} \propto 1/\mathsf{bias}$. In other words, to be able to employ $\hat{h}_{\texttt{MSE}}$ in the first place, one needs to assume the existence of a misspecification error (bias), but it is this very same bias that makes inference invalid when the MSE-optimal point estimator is used for inference purposes.

The invalidity of inference procedures based on the MSE-optimal point estimator was ignored for some time among practitioners. Calonico et al. (2014) highlighted the detrimental consequences of ignoring this misspecification bias and, to solve this inferential problem, proposed a new inference approach based on bias correction of the point estimate, coined *robust bias correction*. The idea behind this method, which allows employing the MSE optimal bandwidth and point estimator, is to adjust the MSE-optimal RD point estimator by estimating its bias and also to adjust the variance estimator used for Studentization purposes when conducting inference. For example, when compared to conventional confidence intervals based on ad-hoc neighborhood selection that rely on standard least squares results, robust bias correction adjusts this confidence interval by recentering (bias correction) and rescaling (robust variance estimator) it. The robust bias corrected RD confidence intervals are fully compatible with employing observations with score lying inside the MSE-optimal selected neighborhood $[\bar{x} - \hat{h}_{\texttt{MSE}}, \bar{x} + \hat{h}_{\texttt{MSE}}]$, while still giving valid inference methods.

Furthermore, Calonico et al. (2016b,a) recently showed that robust bias correction gives demonstrably superior inference when compared to alternative methods employing smaller neighborhoods than $[\bar{x} - \hat{h}_{\texttt{MSE}}, \bar{x} + \hat{h}_{\texttt{MSE}}]$, that is, when shrinking $\hat{h}_{\texttt{MSE}}$ (known as undersmoothing).

In sum, although the MSE-optimal neighborhood can be used for optimal point estimation, standard least squares inference methods cannot be used for inference, and robust bias corrected confidence intervals and related procedures should be used instead. Estimation and robust bias-corrected inference employing the MSE-optimal neighborhood is more robust to the presence of misspecification bias because it does not rely on strong functional

form assumptions about the unknown conditional expectations. As a consequence, these empirical methods are preferred to those relying on ad-hoc neighborhood selectors.

Local Polynomial Neighborhood: CE-Optimal Robust Bias-Corrected Inference

The MSE-optimal neighborhood $[\bar{x} - \hat{h}_{\texttt{MSE}}, \bar{x} + \hat{h}_{\texttt{MSE}}]$ is quite popular in empirical work because it gives an optimal RD treatment effect estimator. As discussed above, the same neighborhood can be used for inference when robust bias correction techniques are employed. However, this neighborhood need not be optimal when the goal is inference. Indeed, Calonico et al. (2016b,a) showed that a different, smaller neighborhood must be used when the goal is constructing optimal confidence intervals in the sense of having the smallest coverage error (CE) probability.

To be more precise, the CE-optimal neighborhood around the RD cutoff is $[\bar{x} - h_{\texttt{CE}}, \bar{x} + h_{\texttt{CE}}]$ with $h_{\texttt{CE}} = C_{\texttt{CE}} \cdot n^{-1/(p+3)}$ and $C_{\texttt{CE}}$ another unknown constant, fundamentally different from $C_{\texttt{MSE}}$, which needs to be estimated in practice because it also involves unknown quantities. This new neighborhood offers robust bias corrected confidence intervals with demonstrably superior optimality properties for inference, when compared to those confidence intervals constructed using the MSE-optimal neighborhood $[\bar{x} - h_{\texttt{MSE}}, \bar{x} + h_{\texttt{MSE}}]$. It follows that $[\bar{x} - h_{\texttt{CE}}, \bar{x} + h_{\texttt{CE}}] \subset [\bar{x} - h_{\texttt{MSE}}, \bar{x} + h_{\texttt{MSE}}]$, in large samples, because $h_{\texttt{CE}} < h_{\texttt{MSE}}$. The same logic also applies to their estimated versions. Figure 1 depicts the CE-optimal choice.

Therefore, in empirical applications, the MSE-optimal neighborhood $[\bar{x} - \hat{h}_{\texttt{MSE}}, \bar{x} + \hat{h}_{\texttt{MSE}}]$ can be used for MSE-optimal RD treatment effect point estimation, and the CE-optimal neighborhood $[\bar{x} - \hat{h}_{\texttt{CE}}, \bar{x} + \hat{h}_{\texttt{CE}}]$, with $\hat{h}_{\texttt{CE}}$ denoting a data-driven implementation of $h_{\texttt{CE}}$, can be used to form CE-optimal robust bias corrected confidence intervals. Employing observations with their score within the CE-optimal neighborhood for point estimation purposes is theoretically allowed but not advisable because the resulting RD treatment effect estimator will have too much variability.

As is the case for the MSE-optimal estimation and robust bias-corrected methods discussed previously, the CE-optimal inference methods are more robust than those based on ad-hoc neighborhood selectors because they optimally trade off misspecification bias underlying the local polynomial approximations to the unknown regression functions, variability of the test statistic (not just the point estimator), and other features of the underlying unknown data generating process.

Local Randomization Neighborhood

The neighborhood selection approaches outlined so far are all related, one way or another, to local or global polynomial regression approximations of the unknown conditional expectations. As such, these methods are based on extrapolation towards the cutoff point $\bar{x}$, using either observations near the cutoff but within the selected neighborhood or simply using all observations in the sample. An alternative approach for identification, estimation and inference in RD designs is based on the idea of local randomization, which assumes that there exists a neighborhood around the cutoff where the underlying data generating process is one (approximately) mimicking a randomized controlled trial (RCT). This heuristic idea

was originally put forward by Lee (2008), and formally developed in Cattaneo et al. (2015), Cattaneo et al. (2017), Sekhon and Titiunik (2017), and references therein.

From this point of view, neighborhood selection is quite different because substantially different assumptions are placed on the underlying data generating process. In other words, none of the neighborhood selectors discussed previously can be used within the local randomization framework because it would be very difficult to rationalize their validity. Cattaneo et al. (2015, 2017) introduced a new neighborhood selection approach: instead of optimizing a point estimator in a mean squared error sense or a confidence interval in a coverage error sense, their idea is to employ pre-intervention covariates and optimize in the sense of minimizing the statistical evidence against the local randomization assumption. To be more precise, the proposal is to conduct a sequence of "balance" or "placebo" tests of no treatment effect on exogenous covariates known to be unaffected by treatment near the RD cutoff, for different proposed neighborhoods, and then select the largest neighborhood that is compatible with local randomization (i.e., the largest neighborhood for which the null hypothesis is not rejected). Under regularity conditions, this method will select a valid neighborhood, which will tend to be smaller than the true neighborhood because no correction for multiple testing is used. Since by construction the neighborhoods are nested, not using multiple testing corrections is appropriate from a robustness perspective in this case.

This neighborhood selection method based on pre-intervention covariate balance tests is similar in spirit to procedures commonly used in the *matching* literature to select a *matched* sample when analyzing observational data under a conditional independence or ignorability assumption (e.g., Imbens and Rubin, 2015). Despite the similarities, the RD local randomization neighborhood selection method is different in that it explicitly exploits the structure of the RD design by localizing near the cutoff and crucially relying on balance tests in a sequence of nested windows.

While the neighborhood selector described above, and subsequent inference procedures, could be implemented via standard large sample estimation and inference methods for RCTs, Cattaneo et al. (2015, 2017) propose to employ randomization inference methods, which are finite sample valid. The main rationale underlying this proposal is at the heart of the specific setting of RD designs: a local randomization assumption in RD designs is most likely to hold, or at least give a good approximation, in a very small neighborhood around the RD cutoff where usually very few observations are available for estimation and inference. Therefore, randomization inference methods, or other analogous finite sample valid methods such as permutation inference, are most useful in the RD context because large sample approximations are unlikely to provide a good enough approximation. Applying the above neighborhood selector to several applications, we have systematically found very small neighborhoods. Thus, based on the methodological arguments and empirical evidence, Figure 1 depicts the local randomization neighborhood as the smallest of all the possible neighborhoods available for estimation and inference in RD designs.

Local randomization methods are fundamentally different from local polynomial methods, both in assumptions and implementation, and therefore they provide a useful robustness check whenever both methods can be used. Furthermore, another important advantage of local randomization methods is that they can handle discrete running variables without any additional assumptions, and randomization inference methods are again most natural whenever the sample size is small. In contrast, local polynomial methods would require

additional parametric assumptions to be valid when the running variable is discrete. This fact is neither surprising nor worrisome, however, since when the running variable is actually discrete there is no need for extrapolation to begin with. It is much more natural and useful to simply consider only the observations having their running variable at the closest discrete value(s) relative to the RD cutoff, on either side, and then use them to conduct estimation and inference. This, of course, changes slightly the parameter of interest, though this is quite natural whenever the running variable has a discrete distribution.

Falsification/Validation Neighborhood

Our discussion so far has focused on neighborhood selection around the RD cutoff for estimation and inference, explicitly relying on different assumptions (i.e., parametric modeling, nonparametric modeling, local randomization). In this subsection, we briefly discuss the related issue of neighborhood selection for falsification/validation of RD designs.

There are two basic falsification/validation methods in the RD literature: (i) tests looking at the continuity of the density of the running variable, and (ii) tests looking at the absence of RD treatment effects on pre-intervention covariates and "placebo" or unaffected outcomes. Both of these approaches also require "localizing" around the RD cutoff. Calonico et al. (2015) discuss related graphical falsification and presentation methods using RD plots, which we do not discuss here to conserve space.

Continuity in the density of the running variable was originally proposed by McCrary (2008), and is by now extremely popular in empirical work. This test is usually understood as providing evidence, or lack thereof, of units having intentionally changed or manipulated their score value near the cutoff. Cattaneo et al. (2016a) recently developed a more robust, nonparametric local polynomial inference method that avoids selecting multiple tuning parameters when implementing this density test. In their approach, the neighborhood is selected in a data-driven, objective way with the explicit goal of minimizing the MSE of the density estimators used to construct the test statistic. It is not possible to determine whether this MSE-optimal neighborhood will be larger or smaller than any of the neighborhoods described previously, because the objective and estimation methods are quite different (i.e., density estimation vs. conditional expectation estimation). What is clear is that the neighborhood for the density test should not be equal, in general, to any of the other neighborhoods: i.e., it should be chosen explicitly for the goal at hand, falsification testing based on local polynomial density estimation. In addition, Frandsen (2017) also developed a "continuity in density" testing approach for the case of discrete running variable. For this method, at present, there is no optimal way of choosing a neighborhood beyond some ad-hoc selection, though the procedure allows for very few "observations" (mass points) near the cutoff because it relies on finite sample inference methods (formally justified by some large sample approximations). Again, there is no reason why the "neighborhood" used for this density test with discrete running variable should coincide with any of the other neighborhoods, and in general it will not.

The density test is quite useful and intuitive because it exploits some of the specific features of RD designs. The second falsification/validation method commonly used in practice is more standard, in the sense that it is directly imported from common practice in other experimental and non-experimental settings. Specifically, this second method seeks to test

whether there is evidence of an RD treatment effect on covariates and outcomes that should (or, at least, are assumed to) be unaffected by the treatment. This approach is conceptually analogous to testing for a treatment effect on pre-intervention covariates in the context of RCTs, and can be implemented using directly the modern local polynomial and randomization inference methods described in the previous sections for RD estimation and inference. As an alternative, Canay and Kamat (2016) have recently proposed a permutation inference approach for falsification testing based on comparing the whole distribution of treatment and control groups, which is also justified via some large sample approximations near the cutoff. The authors conduct neighborhood selection using a rule-of-thumb based on a simple reference model, which leads to yet another neighborhood to be used in applications when implementing their method.

Recommendations for Practice and Final Remarks

To conclude, we offer some practical recommendations for empirical work. We build on the methodological points put forward above, and hence only offer very brief takeaway methodological points:

1. Always employ RD optimal data-driven neighborhood (bandwidth or window) selectors, at least as a benchmark or starting point. This gives objectivity and robustness because it incorporates explicitly empirical features such as density of observations, variability of the data, or curvature of the unknown regression functions, in a principled way.

2. Employ data-driven neighborhood (bandwidth or window) selectors according to the specific goal and assumptions imposed, which should also be explicitly stated and explained. There is no one neighborhood selector appropriate for all objectives when using local polynomial approximations, and even for local randomization methods sensitivity analysis with respect to the neighborhood used is very important.

3. Do not employ the *same* neighborhood for different outcome variables, pre-intervention covariates (if conducting falsification testing), estimation and inference procedures, or falsification methods. Using the same neighborhood for different goals, outcomes or samples disregards the specific empirical features (e.g., number of observations near the cutoff, variability or curvature), and will lead to unreliable empirical results due to invalidity of the methods employed.

Thistlethwaite and Campbell (1960) introduced one of the best non-experimental methods for the analysis and interpretation of observational studies. In recent years many methodological and theoretical developments not only have extended the basic regression discontinuity design to many other settings, but also have provided major improvements in terms of presentation, estimation, inference and falsification for empirical practice. In this discussion, we focused on arguably the most important and challenging part of analyzing and implementing RD designs: neighborhood, bandwidth or window selection around the RD cutoff. Much methodological progress has been achieved in recent years regarding this important task, making RD designs even more credible and robust in applications.

Acknowledgments

We thank our close collaborators and colleagues, Sebastian Calonico, Max Farrell, Michael Jansson, Xinwei Ma, and Rocio Titiunik, whose ideas and criticisms over the years have shaped this discussion. We also thank Justin McCrary and David McKenzie for recent energizing discussions and excellent comments on RD design methodology. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grant SES-1357561.

`R` and `Stata` software packages implementing the main neighborhood (i.e., bandwidth or window) selectors discussed above are available at:

https://sites.google.com/site/rdpackages

References

Bartalotti, O. and Brummet, Q. (2017). Regression discontinuity designs with clustered data. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, *forthcoming*.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016a). Coverage error optimal confidence intervals for regression discontinuity designs. Working paper, University of Michigan.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016b). On the effect of bias estimation on coverage accuracy in nonparametric inference. arXiv:1508.02973.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2016c). Regression discontinuity designs using covariates. Working paper, University of Michigan.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, 110(512):1753–1769.

Canay, I. A. and Kamat, V. (2016). Approximate permutation tests and induced order statistics in the regression discontinuity design. Working paper, Northwestern University.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2017). Regression kink design: Theory and practice. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, *forthcoming*.

Cattaneo, M. D. and Escanciano, J. C. (2017). *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, forthcoming.

Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. D., Jansson, M., and Ma, X. (2016a). Simple local regression distribution estimators with an application to manipulation testing. Working paper, University of Michigan.

Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2016b). Interpreting regression discontinuity designs with multiple cutoffs. *Journal of Politics*, 78(4):1229–1248.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017). Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, forthcoming.

Cerulli, G., Dong, Y., Lewbel, A., and Poulsen, A. (2017). Testing stability of regression discontinuity models. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing*, forthcoming*.

Cook, T. D. (2008). "waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654.

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

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC, New York.

Frandsen, B. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing*, forthcoming*.

Gelman, A. and Imbens, G. W. (2014). Why high-order polynomials should not be used in regression discontinuity designs. NBER working paper 20405.

Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Imbens, G. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jales, H. and Yu, Z. (2017). Identification and estimation using a density discontinuity approach. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, *forthcoming*.

Keele, L., Lorch, S., Passarella, M., Small, D., and Titiunik, R. (2017). An overview of geographically discontinuous treatment assignments with an application to children's health insurance. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, *forthcoming*.

Keele, L. J. and Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155.

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Sekhon, J. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, *forthcoming*.

Shen, S. and Zhang, X. (2016). Distributional regression discontinuity: Theory and applications. *Review of Economics and Statistics*, forthcoming.

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.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Chapman & Hall/CRC, Florida.

Observational Studies 2 (2016) 147-155 Submitted 11/16; Published 12/16

Regression Discontinuity Designs in the Econometrics Literature

Guido W. Imbens **imbens@stanford.edu**
Graduate School of Business, Stanford University, SIEPR, and NBER
Stanford, CA, U.S.A.

Abstract

Many decades after being introduced by Thistlewaite and Campbell (1960), regression discontinuity designs have become an important tool for causal inference in social sciences. Researchers have found the methods to be widely applicable in settings where eligibility or incentives for participation in programs is at least partially regulated. Alongside, and motivated by, the many studies applying regression discontinuity methods there have been a number of methodological studies improving our understanding, and implementation, of, these methods. Here I report on some of the recent advances in the econometrics literature.

Keywords: regression discontinuity designs, matching, regression kink designs, local linear regression

1. Introduction

In the late 1990s and early 2000s, regression discontinuity designs (rdd's for short), originally developed many years earlier by Thistlewaite and Campbell (1960), enjoyed a renaissance in social science in general, and in economics in particular. As the rdd method became one of the most popular strategies for identifying causal effects (Angrist and Pischke (2008)) and a standard topic in first year econometrics courses in PhD programs, researchers became increasingly aware of the wide applicability of the methods developed by Thistlewaite and Campbell. Early applications in economics include Black (1999), using geographical boundaries, Van Der Klaauw (2002), using college application thresholds, and Lee (2008), using election thresholds. See Cook (2008) for a historical perspective, including references to earlier discussions in economics that failed to catch on, and for recent general discussions and surveys in the economics literature see Imbens and Lemieux (2008); Van Der Klaauw (2008); Lee and Lemieux (2010); Calonico et al. (2015); Choi and Lee (2016). For general discussions outside of economics see Trochim (1984); Shadish et al. (2002); Skovron and Titiunik (2015). The recent increase in applications in economics has motivated novel theoretical work on rdd methods in econometrics that have improved our understanding of rdd methods, as well as affected empirical practice. Here I want to discuss some of these recent methodological innovations.

2. Basic Set Up

As is common in the econometric causal literature, though not originally in the rdd literature, we set the problem up in the Rubin Causal Model or potential outcome framework

(Rubin (1974); Holland (1986); Imbens and Rubin (2015)). We assume there are, for each unit in a large population, two potential outcomes, $Y_i(0)$ and $Y_i(1)$ for unit i, corresponding to the control and treated outcome, with the unit-level causal effect some comparison of the two, e.g., the difference $Y_i(1) - Y_i(0)$. There is a binary treatment $W_i \in \{0, 1\}$, defining the observed outcome $Y_i^{\rm obs} = Y_i(W_i)$, and an exogenous forcing variable X_i, as well as possibly additional covariates Z_i. At the threshold, say $x = 0$, the probability of receiving the treatment changes discontinuously. If it changes from zero to one we have a sharp rd design, otherwise a fuzzy rd design. In general the estimand in rdd analyses is the ratio of two discontinuities, first, the discontinuity in the conditional expectation of the realized outcome given the forcing variable, and second, the treatment indicator given the forcing variable, both at the threshold:

$$\tau = \frac{\lim_{x\downarrow 0} \mathbb{E}[Y_i^{\rm obs}|X_i = x] - \lim_{x\uparrow 0} \mathbb{E}[Y_i^{\rm obs}|X_i = x]}{\lim_{x\downarrow 0} \mathbb{E}[W_i|X_i = x] - \lim_{x\uparrow 0} \mathbb{E}[W_i|X_i = x]}.$$

In the sharp rdd setting the denominator is excactly one and we simply look at the magnitude of the discontinuity in the conditional expectation of the outcome at the threshold.

3. The Interpretation of Fuzzy Regression Discontinuity Designs

The first, and arguably most important, innovation in the econometrics literature concerns the precise interpretation of the estimand in fuzzy rd designs in settings with heterogenous treatment effects. Although fuzzy regression discontinuity designs had been around at least since Trochim (1984), their analysis was limited to the case with constant treatment effects. Hahn et al. (2001) (HTV from hereon) established an important link to the instrumental variables literature with heterogenous treatment effects. In particular they showed that in the fuzzy rdd the ratio of discontinuities in the conditional mean for the outcome and the conditional mean for the treatment has an interpretation of a local average treatment effect (Imbens and Angrist (1994); Angrist et al. (1996)). The HTV argument shows that the fuzzy rdd estimand is the average effect of the treatment, for the subpopulation of compliers, among those with a value for the forcing variable close to the threshold. Compliers in this subpopulation of individuals with a value for the forcing variable close to the threshold are individuals for whom it matters which side of the threshold they are on. The argument is slightly subtle, because it relies on a clear interpretation of what is stochastic in this setting. Specifically, for an individual close to, but on the left of the threshold, say with $X_i \in (-\epsilon, 0)$, it requires one to think about what would have happened to this individual had they been on the other side of the threshold. We can do this in two ways. In the HTV approach, the forcing variable X_i is taken as potentially manipulable, so that one can think of potential treatment values $W_i^{\rm htv}(x)$ for different values of the forcing variable, with $W_i = W_i^{\rm htv}(X_i)$ the realized value. Compliers are in this approach individuals who would have received the treatment had X_i been slightly above the threshold, but not if X_i had been slightly below the threshold:

$$C_i^{\rm htv} = \begin{cases} a & \text{if } \lim_{x\downarrow 0} W_i^{\rm htv}(x) = 1,\ \lim_{x\uparrow 0} W_i^{\rm htv}(x) = 1, \\ c & \text{if } \lim_{c\downarrow 0} W_i^{\rm htv}(x) = 0,\ \lim_{c\uparrow 0} W_i^{\rm htv}(x) = 1, \\ n & \text{if } \lim_{c\downarrow 0} W_i^{\rm htv}(x) = 0,\ \lim_{c\uparrow 0} W_i^{\rm htv}(x) = 0, \\ d & \text{if } \lim_{c\downarrow 0} W_i^{\rm htv}(x) = 1,\ \lim_{c\uparrow 0} W_i^{\rm htv}(x) = 0. \end{cases}$$

Typically the presence of defiers is ruled out. In this perspective the forcing variable is stochastic, rather than a fixed characteristic of the individual, and could have taken on a different value for a given individual from that observed one for that individual.

In some applications it may be difficult to imagine the forcing variable as a causal variable, defining potential outcomes, say in the case where the forcing variable is a fixed immutable characteristic such age. In such cases an alternative, following Bertanha and Imbens (2016), may be to view the threshold, rather than the forcing variable, as manipulable, generating potential treatment values corresponding to the threshold: $W_i^{\rm bi}(c)$ is in this approach the treatment level for unit i if the threshold were set at c, where we only consider values of c close to actual threshold of zero. In this perspective compliers are defined by the pair of limits of $W_i^{\rm bi}(c)$, taken from the left and from the right of the actual threshold value zero:

$$C_i^{\rm bi} = \begin{cases} a & \text{if } \lim_{c\downarrow 0} W_i^{\rm bi}(c) = 1,\ \lim_{c\uparrow 0} W_i^{\rm bi}(c) = 1, \\ c & \text{if } \lim_{c\downarrow 0} W_i^{\rm bi}(c) = 0,\ \lim_{c\uparrow 0} W_i^{\rm bi}(c) = 1, \\ n & \text{if } \lim_{c\downarrow 0} W_i^{\rm bi}(c) = 0,\ \lim_{c\uparrow 0} W_i^{\rm bi}(c) = 0, \\ d & \text{if } \lim_{c\downarrow 0} W_i^{\rm bi}(c) = 1,\ \lim_{c\uparrow 0} W_i^{\rm bi}(c) = 0. \end{cases}$$

Again we typically rule out the presence of defiers.

This difference in interpretation has some conceptual implications. If one views the forcing variable as stochastic, it can be used to generate a randomization distribution for the regression discontinuity estimator with approximately independent treatment assignments, similar to that in a randomized experiment. Using only individuals close to the threshold, we have essentially a randomized experiment with assignment for all units close to independent. However, if we view the threshold as potentially manipulable, there is only a single stochastic component driving the randomization properties of the estimator, so that the treatment assigments are closely related, and the fundamental difference with an actual randomized experiment becomes clear.

4. Supplementary Analyses in Regression Discontinuity Designs

A second active area of methodological innovations has been the use of, what Athey and Imbens (2016) call in general discussion of causal inference, supplementary analyses. They define these as analyses where the aim is not to get a better estimate of the object of interest, that is the causal effect. Rather they are analyses that are intended to provide support for the main analyses, by disproving potential arguments against the validity of the main analyses. Depending on the results of the supplementary analyses the credibility of the main analyses is either weakened or strengthened.

One of the major concerns in rdd analyses is that the forcing variable may have been manipulated. In many cases there are substantial costs or benefits for the agents associated with being just to the left or right of the threshold associated with the change in incentives. If agents have some ability to change the actual, or even just the recorded, value of the forcing variable, they would in that case have a strong incentive to do so. A classic example is that of tests scores used to decide on student's eligibility of attractive educational options, or to decide on required remedial programs. If there is discretion in the grading of the tests, and the graders are aware of both the importance of the test, and of the value of the

threshold, and if the graders have preferences over the outcomes for the students, they may change grades for some individuals sufficiently close to the threshold.

It can be challenging to address this type of manipulation through the statistical analysis, although there are some interesting approaches involving shape restrictions on the underlying potential outcome distributions (Diamond and Persson (2016)). Much easier is the task of establishing whether such manipulation is taking place. If there is, one would expect a discontinuity in the marginal density of the forcing variable because for individuals on one side of the threshold there, and for individuals on the other side of the threshold there is no, incentive to manipulate the score. McCrary (2008) developed a test for the null hypothesis of no discontinuity in the density of the forcing variable that should be performed any time someone does a rdd analysis. See also Otsu et al. (2015) for an alternative version of the test. Note that, for the purpose of estimating the difference in the conditional means of the outcome on the right and the left of the threshold, there is formally no need for the marginal density of the forcing variable to be continuous at that point. The reason that the test is important is that the argument that underlies the identification strategy, and in particular the notion that individuals directly to the left and the right of the threshold are comparable other than through the receipt of the treatment, is difficult to reconcile with finding that there are substantially fewer people just to the right than to the left of the threshold.

A second supplementary analyses in the rdd setting involves checking the continuity of the conditional expectation of exogenous variables around the threshold. Again this continuity is not required for consistency, but a discontinuity in such conditional expectations is difficult to reconcile with comparability of the individuals to the left and the right of the threshold, and would suggest the possibility of unobserved imbalances on the right and the left. Such analyses are similar to placebo analyses in studies of causal effects under unconfoundedness, where often tests for zero effects on lagged outcomes are presented to assess unconfoundedness (Athey and Imbens (2016); Imbens and Rubin (2015)).

5. Bandwidth Choice in Regression Discontinuity Designs

The currently preferred analysis in rdd settings, e.g., Hahn et al. (2001); Porter (2003) is to use local linear, or sometimes local quadratic methods (Calonico et al. (2014)) rather than simple kernel estimators or global high order polynomial methods. Simple kernel regression methods have poor properties at the boundary of the support, and that is precisely where we are interested in the estimates in this setting. Gelman and Imbens (2014) argue against the use of global high-order polynomials because of poor properties in terms of mean-squared error, coverage rates for confidence intervals and the difficulties in selecting the order of the polynomial. Given the use of local regression methods, the question is how to choose the degree of localness, that is, the bandwidth. Early on in the literature common practice was to use off-the-shelf bandwidth selection methods based on crossvalidation, e.g., Ludwig and Miller (2005). However, crossvalidation methods are not as attractive here as they are for bandwidth selection in nonparametric regression because in the current setting we are interested in the value of the regression only at a few points. More recently bandwidth selection methods have been developed that are specifically geared towards the goal of precisely estimating the magnitude of the discontinuity, at the threshold, in the conditional

expectation of the outcome given the forcing variable, (Imbens and Kalyanaraman (2012); Calonico et al. (2014)).

These bandwidth selection methods are based on asymptotically balancing the square of the bias and the variance of the estimator for the limit of the value of the regression function at the threshold, from the right and the left.

6. External Validity in Regression Discontinuity Designs

One, and perhaps the, major limitation of rdd analyses is the lack of external validity. In many cases the methods lead to estimates with a high degree of internal validity, but the conclusions are limited in two aspects. First, they are restricted to the subpopulation of compliers, and second they are restricted to individuals with values for the forcing variable close to the threshold. Recently there has been some work examining the presence and credibility of any evidence that these estimates have wider external validity, be that for non-compliers, or for units with values of the forcing variable away from the threshold.

First consider only units close to the threshold. Battistin and Rettore (2008) and Bertanha and Imbens (2016) propose comparing compliers without the treatment ("control compliers") to never-takers at the threshold, and comparing compliers with the treatment ("treated compliers") to always-takers at the threshold. If one clearly rejects the null hypotheses that, say never-takers and control compliers, are comparable, than it appears less plausible that the average treatment effect for compliers (which is estimable) is useful as a predictor for the average effect of the treatment for never-takers (which we cannot estimate directly). If, on the other hand, we find that treated compliers are comparable to always-takers, and control compliers are comparable to never-takers, it appears more plausible that the average effect for compliers is indicative of the average effect for the other subpopulations close to the threshold. In that case the external validity of the rdd estimates is enhanced. Note that the same argument can be made in for other, non-rdd versions of instrumental variables, and it is related to the discussion on testing in Angrist (2004). Bertanha and Imbens (2016) point out that there is a convenient graphical interpretation of this null hypothesis, namely the continuity of the conditional expectation of the outcome as a function of the forcing variable, conditional on the treatment group, adjusted for other covariates.

Angrist and Fernandez-Val (2010) and Angrist and Rokkanen (2015) take different approaches to the extrapolation to other subpopulations. In the context of instrumental variables estimation, but in a way that can conceptually easily be extended to rdd settings, Angrist and Fernandez-Val (2010) focus on the difference between estimators based on unconfoundedness and estimators based on iv or rdd assumptions (in both sharp and fuzzy rdd settings). If exogenous covariates can eliminate the differences between the two, they argue that extrapolating the complier effects to the general population is more plausible. In the context of sharp rd designs Angrist and Rokkanen (2015) focus on the role of covariates to eliminate differences between units with different values of the forcing variable. If the other covariates can eliminate all or most of the association between the forcing variable and the outcomes away from the threshold, again it becomes more plausible to extrapolate the estimated effects at the threshold to other subpopulations.

Dong and Lewbel (2015) point out that under the rdd assumptions one can in principle identify not simply the level of the conditional expectation on both sides of the threshold, but also derivatives of this conditional expectation. They explore using estimates of these derivatives to extrapolate away from the threshold.

7. Multiple Thresholds and Multiple Assignment Variables

In many applications the assignment process is more complex than covered by the simple rdd setting. There may be multiple thresholds at which incentives to participate in a program change discontinuously, as in Bertanha (2015); Abdulkadiroğlu et al. (2014). In many such cases there is not sufficient information at a single threshold to obtain precise estimates of the causal effect at that threshold. In that case one may wish to combine the estimates at the different thresholds into a single average effect.

There may also be multiple measures that enter into the eligibility decision, as in Papay et al. (2011); Imbens and Zajonc (2011). For example, a student may be required to participate in a remedial program unless the student receives a passing grade in both mathematics and reading tests. In this case the researcher has several options. One can analyze the data using the minimum of the two grades in a sharp rdd. In that case one can also assess heterogeneity in the effects by comparing individuals close to the reading threshold among the subpopulation with mathematics test scores above the threshold, or the other way around. One can also analyze the data using either the reading or mathematics score as a forcing variable in a fuzzy rdd.

8. Regression Kink Designs

A very recent generalization of regression discontinuity designs is what has been labelled the regression kink design (rkd), Card et al. (2015) and Dong (2014). In this case the treatment of interest is a continuous one. At the threshold the conditional expectation of the outcome is not expected to change discontinuously. However, the derivative of the conditional expectation at that point may change discontinuously, leading to a kink in the conditional expectation, lending the approach its name. The discontinuity in the derivative of the conditional expectation of the outcome is attributed to the the discontinuity in the derivative of the conditional expectation of the treatment given the forcing variable changes at the threshold. For example, consider a case where the researcher is interested in the effect of unemployment benefits on subsequent earnings. The treatment of interest is the benefit level an individual receives. The forcing variable may be prior earnings, in a setting where the benefits decrease with earnings, with the rate of decrease changing discontinuously at the threshold. Card et al. (2015) and Dong (2014) extend rdd methods to such cases. Obviously estimating the change in the derivatives is a more challenging task than estimating the change in the level of a conditional expectation, and consequently regression kink analyses will require in practice more data than regression discontinuity analyses.

9. Conclusion

In this note I discuss some of the recent work in econometrics on regression discontinuity designs. Decades after these methods were first introduced by Thistlewaite and Campbell (1960), they are now among the most widely used methods for causal inference in economics and other social sciences. This has motivated more methodological advances in what is currently a very active research area.

References

Abdulkadiroğlu, A., Angrist, J., and Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, 82(1):137–196.

Angrist, J. and Fernandez-Val, I. (2010). Extrapolate-ing: External validity and overidentification in the late framework. Technical report, National Bureau of Economic Research.

Angrist, J. and Pischke, S. (2008). *Mostly Harmless Econometrics: An Empiricists' Companion*. Princeton University Press.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–472.

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

Athey, S. and Imbens, G. (2016). The state of applied econometrics-causality and policy evaluation. *arXiv preprint arXiv:1607.00699*.

Battistin, E. and Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, 142(2):715–730.

Bertanha, M. (2015). Regression discontinuity design with many thresholds. *Available at SSRN*.

Bertanha, M. and Imbens, G. (2016). External validity in fuzzy regression discontinuity designs. CORE Discussion Paper 2016/25.

Black, S. (1999). Do better schools matter? parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599.

Calonico, S., Cattaneo, M., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Calonico, S., Cattaneo, M., and Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.

Choi, J.-y. and Lee, M.-j. (2016). Regression discontinuity: review with extensions. *Statistical Papers*, pages 1–30.

Cook, T. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654.

Diamond, R. and Persson, P. (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Technical report, National Bureau of Economic Research.

Dong, Y. (2014). Jump or kink? identification of binary treatment regression discontinuity design without the discontinuity. *Unpublished manuscript*.

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

Gelman, A. and Imbens, G. (2014). Why high-order polynomials should not be used in regression discontinuity designs. NBER Working Paper No. 20405.

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

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–970.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Imbens, G. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Imbens, G. and Zajonc, T. (2011). Regression discontinuity design with multiple forcing variables. *Unpublished manuscript*.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 61:467–476.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Lee, D. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355.

Ludwig, J. and Miller, D. L. (2005). Does head start improve children's life chances? evidence from a regression discontinuity design. Technical report, National Bureau of Economic Research.

McCrary, J. (2008). Testing for manipulation of the running variable in the regression discontinuity design. *Journal of Econometrics*, 142(2):698–714.

Otsu, T., Xu, K.-L., and Matsushita, Y. (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics*, 186(1):94–112.

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161(2):203–207.

Porter, J. (2003). Estimation in the regression discontinuity model. Available on CiteSeer.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company.

Skovron, C. and Titiunik, R. (2015). A practical guide to regression discontinuity designs in political science. Technical report, working paper, University of Michigan.

Thistlewaite, D. and Campbell, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(2):309–317.

Trochim, W. M. (1984). *Research design for program evaluation: The regression-discontinuity approach*, volume 6. SAGE Publications, Inc.

Van Der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43(2):1249–1287.

Van Der Klaauw, W. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour*, 22(2):219–245.

Observational Studies 2 (2016) 156-173 Submitted 10/16; Published 12/16

Regression Discontinuity Designs as Local Randomized Experiments

Alessandra Mattei **mattei@disia.unifi.it**
Department of Statistica, Informatica, Applicazioni
University of Florence
Viale Morgagni 59, 50134 Firenze, Italy

Fabrizia Mealli **mealli@disia.unifi.it**
Department of Statistica, Informatica, Applicazioni
University of Florence
Viale Morgagni 59, 50134 Firenze, Italy

Abstract

In the seminal paper from 1960, Thistlethwaite and Campbell (1960) introduce the key ideas underlying regression discontinuity (RD) designs, which, even if initially almost completely ignored, have then acted as a fuse of a blowing number of studies applying and extending RD designs starting from the late nineties. Building on the original idea by Thistlethwaite and Campbell (1960), RD designs have been often described as designs that lead to locally randomized experiments for units with a realized value of a so-called forcing variable falling around a pre-fixed threshold. We embrace this perspective, and in this discussion we offer our view on how the original proposal by Thistlethwaite and Campbell (1960) should be formalized. We introduce an explicit local overlap assumption for a subpopulation around the threshold, for which we re-formulate the Stable Unit Treatment Value Assumption (SUTVA), and provide a formal definition of the hypothetical experiment underlying RD designs, by invoking a local randomization assumption. A distinguishing feature of this approach is that it embeds RD designs in a framework that is fully consistent with the potential outcome approach to causal inference. We discuss how to select suitable subpopulation(s) around the threshold with adjustment for multiple comparisons, and how to draw inference for the causal estimands of interest in this framework. We illustrate our approach in a study concerning the effects of University grants on students' dropout.

Keywords: Causal Inference, Local Causal Effects, Local Randomization, Potential Outcomes, Regression Discontinuity Designs

1. Introduction

Thistlethwaite and Campbell (1960) are considered to be the fathers of the regression discontinuity (RD) design and they deserve great recognition for their outstanding insight. It is a pleasure and an honor for us to contribute to the discussion on the reprint of their original article on the RD design.

In RD designs, the assignment to the treatment is determined, at least partly, by the realized value of a variable, usually called the forcing or running variable, falling on either side of a prefixed threshold or cutoff point. Thistlewaite and Campbell's key intuition

was that in RD designs the comparisons of units with very close values of the forcing variable, namely around the point where the discontinuity is observed, but different levels of treatment, may lead to valid inference on causal effects of the treatment at the threshold. Nevertheless Thistlethwaite and Campbell (1960) provided no formal description of the design and no theoretical result. In practice, the approach they proposed was a regression analysis with a causal interpretation, and indeed they referred to it as a "RD analysis", rather than a "RD design." It was only later that Campbell (1969) called that type of analysis "a design," but again without giving any formal statistical presentation but only relying on intuitions and analogies to the Fisher's work on design.

Despite Thistlewaite and Campbell's brilliant intuition, RD designs did not attract much attention in the causal inference literature until recently, as the historical excursus in Cook (2008) describes. It is only starting from the late 1990s that RD designs have become increasingly popular in statistics, social science, economics and, more recently also in epidemiology and the medical sciences. In the last two decades, causal inference in RD designs has been a fertile area of research, and there has been a growing number of studies applying and extending RD methods. General surveys can be found in Imbens and Lemieux (2008) and Lee and Lemieux (2010). See also Athey and Imbens (2016) and the edited volume by Cattaneo and Escanciano (2016) for more recent reviews, discussions, and references.

In the modern causal inference literature, inference on causal effects in RD designs uses a formal approach to causal inference rather than the regression framework that was originally used by Thistlethwaite and Campbell (1960). Following one of the main strand of the literature, we will frame RD designs in the context of the potential outcome approach to causal inference (Rubin, 1974; Imbens and Rubin, 2015). See Constantinou and O'Keeffe (2016) for an alternative perspective embedded in the decision theoretic approach to causal inference (Dawid, 2000).

Traditionally, the forcing variable in RD settings is viewed as a pretreatment covariate and RD designs are usually described as quasi-experimental designs with a non-probabilistic assignment mechanism. Therefore inference in RD designs needs to rely on some kind of extrapolation: the traditional inference approach in RD designs invokes smoothness assumptions for the relationship between the outcome and the forcing variable, such as continuity of conditional regression functions (or conditional distribution functions) of the outcomes given the forcing variable. Under these smoothness assumptions, which imply randomization at the single threshold value (Battistin and Rettore, 2008), observations near the known cutoff are used to derive estimates of treatment effects *at the threshold*, using global polynomial series estimators or local-polynomial (non-)parametric regression methods and their asymptotic proprieties. In real applications, large-sample approximations might be unreliable, especially if the sample size around the threshold is small, and exact inference might be preferable. Some further discussion on this traditional approach and its implication for inference is offered in Section 5.

Building on the original idea by Thistlethwaite and Campbell (1960), RD designs have been often described as designs that lead to locally randomized experiments around the threshold (Lee, 2008; Lee and Lemieux, 2010; Dinardo and Lee, 2011). Expanding on this interpretation, a recent strand of the literature (e.g., Cattaneo et al., 2015; Li et al., 2015; Sales and Hansen, 2015) is moving towards a formal and well-structured definition of the con-

ditions under which RD designs can be formally described as local randomized experiments, also discussing the relationship between local randomization and smoothness/continuity RD assumptions (de la Cuesta and Imai, 2016; Skovron and Titiunik, 2015). We embrace this new perspective, to which we have also proudly contributed (Li et al., 2015).

In this discussion we offer our view on how the original proposal by Thistlethwaite and Campbell (1960) should be formalized, that is, how their heuristic reasoning can be formally described. Our view is based on the approach we propose in Li et al. (2015). A distinguishing feature of this approach is that it embeds RD designs in a framework that is fully consistent with the potential outcome approach to causal inference, providing a formal definition of the hypothetical experiment underlying RD designs, based on a description of the assignment mechanism, formalized as a unit-exchangeable stochastic function of covariates and potential outcomes.

We provide a detailed description of this approach, discussing both theoretical and practical issues, and highlighting issues that we feel are valuable topics for further research. We focus on the sharp RD design, the original form of the design, where the treatment status is assumed to be a deterministic step function of the forcing variable: All units with a realized value of the forcing variable on one side of a prefixed threshold are assigned to one treatment regime and all units on the other side are assigned to the other regime. Nevertheless, our methodological framework applies also to fuzzy RD designs, where the realized value of the forcing variable does not alone determine the receipt of the treatment, although a value of the forcing variable falling above or below the threshold acts as an encouragement or incentive to participate in the treatment (see Li et al., 2015, for details on the probabilistic formulation of the assignment mechanism underlying fuzzy RD designs).

2. Our Interpretation of RD Designs as Local Randomized Experiments

Consider a sample or population of N units indexed by $i = 1 \ldots, N$. Let S_i denote the forcing variable, on the basis of which a binary treatment Z_i is assigned according to a RD rule: If a unit has a value of S falling below (or above, depending on the specific application) a predetermined threshold, s_0, that unit is assigned to the active treatment, and s/he is assigned to the control treatment otherwise. Therefore the treatment status Z_i for each unit i is a deterministic function of S_i: $Z_i = \mathbf{1}\{S_i \leq s_0\}$ where $\mathbf{1}\{\cdot\}$ is the indicator function.

Thistlethwaite and Campbell describes the approach they propose arguing that

> The argument [justifying a RD analysis] – and the limitations on generality of the result – can be made more specific by considering a "true" experiment for which the regression-discontinuity analysis may be regarded as a substitute. ... a group of commended students who narrowly missed receiving the higher award might be given opportunity of receiving extra recognition. Thus students in Interval 10 in Figure 1 [in a neighborhood of the threshold, s_0] might be randomly assigned to the different treatment of C of M award and no C of M award (Thistlethwaite and Campbell, 1960, page 310).

We propose to formalize their argument, formally reconstructing the hypothetical "true" experiment underlying a RD design using a framework that is fully consistent with the potential outcome approach. Throughout our discussion we also highlight the key differences

between our approach and both the standard approach to RD designs, where smoothness assumptions are invoked to estimate causal effects at the threshold, as well as alternative, more recent, attempts aiming at formally describing RD designs as local randomized experiments.

Our reconstruction starts from re-defining RD designs step-by-step using the potential outcome approach, which has two essential parts: (a) the definition of the primitive concepts – units, treatments and potential outcomes; and (b) the definition of an assignment mechanism determining which potential outcomes are realized, and possibly observed. Formally, the assignment mechanism is a probabilistic model for the assignment variable as a function of potential outcomes and covariates. The careful implementation of these steps is absolutely essential for drawing objective inferences on causal effects in any study, and thus also in RD designs.

In RD designs, the treatment status, which a unit may be exposed to, depends on the forcing variable, which is the assignment variable. Potential outcomes need to be defined accounting for the alternative levels of the forcing variable and the assignment mechanism needs to be specified as probabilistic model for the conditional probability of the forcing variable given potential outcomes and covariates.

In the literature, the forcing variable is traditionally viewed as a pretreatment covariate and RD designs are typically described as designs with an irregular assignment mechanism breaching the overlap assumption: $Pr(Z_i = 1) = Pr(\mathbf{1}\{S_i \leq s_0\})$ and $Pr(\mathbf{1}\{S_i \leq s_0\}) = \mathbf{1}\{S_i \leq s_0\}$, if S is a fixed pretreatment covariate, and thus the probability of assignment to treatment versus control is equal to zero or one for all units.

We revisit this perspective viewing the forcing variable, S, as a random variable with a probability distribution, and propose to break the longtime interpretation of RD designs as an *extreme* violation of the overlap assumption. Specifically, we formulate the following assumption:

Assumption 1 *(Local overlap). Let $\mathcal{U}$ be the random sample (or population) of units in the study. There exists a subset of units, $\mathcal{U}_{s_0}$, such that for each $i \in \mathcal{U}_{s_0}$, $Pr(S_i \leq s_0) > \epsilon$ and $Pr(S_i > s_0) > \epsilon$ for some sufficiently large $\epsilon > 0$.*

Assumption 1 is essentially a local overlap assumption implying that there exists a subpopulation of units, each of whom has a probability of having a value of the forcing variable falling on both sides of the threshold sufficiently faraway from both zero and one. Assumption 1 implies that each unit belonging to a subpopulation $\mathcal{U}_{s_0}$ has a non-zero marginal probability of being assigned to either treatment levels: $0 < Pr(Z_i = 1) < 1$ for all $i \in \mathcal{U}_{s_0}$. Therefore for units belonging to the subpopulation $\mathcal{U}_{s_0}$, an overlap assumption holds, and this represents a main and key distinction with the traditional description of RD designs. Assumption 1 is a *local* overlap assumption in the sense that a unit with a realized value of the forcing variable falling very faraway from the threshold does not probably belong to the subpopulation $\mathcal{U}_{s_0}$ and may have a zero probability of having a value of the forcing value falling on the other side of the threshold.

It is worth noting that Assumption 1 does not require that the subpopulation $\mathcal{U}_{s_0}$ is unique; it only requires that there exists at least one subpopulation $\mathcal{U}_{s_0}$. Also the value ϵ in Assumption 1 has not a substantive meaning, but it is only a methodological tool for formally describing the subpopulation $\mathcal{U}_{s_0}$.

Assumption 1 plays a key role in the definition of the causal estimands: Under Assumption 1, we can focus on causal effects for a subpopulation, $\mathcal{U}_{s_0}$, rather than on causal effects at the threshold, which are the causal estimands typically considered in RD designs. The correct definition of causal effects depends on the specification of potential outcomes. Each unit in the subpopulation $\mathcal{U}_{s_0}$ can be exposed to alternative values of the forcing variable, therefore, in principle, potential outcomes need to be defined as function of the forcing variable. Let $N_{\mathcal{U}_{s_0}}$ be the number of units belonging to a subpopulation $\mathcal{U}_{s_0}$ and let $\mathbf{s}$ be an $N_{\mathcal{U}_{s_0}}-$dimensional vector of values of the forcing variables with ith element s_i. For each unit $i \in \mathcal{U}_{s_0}$, let $Y_i(\mathbf{s})$ denote the potential outcomes for an outcome variable Y: $Y_i(\mathbf{s})$ is the value of Y for unit i given the vector of values of the forcing variable, $\mathbf{s}$.

Working with the potential outcomes $Y_i(\mathbf{s})$ raises serious challenges to causal inference because the forcing variable is a continuous variable, and so generates a continuum of potential outcomes, and potential outcomes for a unit may be affected by the value of the forcing variable of other units. To face these challenges, within the subpopulation $\mathcal{U}_{s_0}$, we formulate a modified Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980), specific to RD settings:

Assumption 2 *(Local RD-SUTVA). For each $i \in \mathcal{U}_{s_0}$, consider two treatment statuses $z_i^{'} = \mathbf{1}(s_i^{'} \leq s_0)$ and $z_i^{''} = \mathbf{1}(s_i^{''} \leq s_0)$, with possibly $s_i^{'} \neq s_i^{''}$. If $z_i^{'} = z_i^{''}$, that is, if either $s_i^{'} \leq s_0$ and $s_i^{''} \leq s_0$, or $s_i^{'} > s_0$ and $s_i^{''} > s_0$, then $Y_i(\mathbf{s}^{'}) = Y_i(\mathbf{s}^{''})$.*

Assumption 2 introduces two important simplifications. First, it rules out interference between units, implying that potential outcomes for a unit cannot be affected by the value of the forcing variable (and by the treatment status) of other units. Second, Local RD-SUTVA implies that for units in the subpopulation $\mathcal{U}_{s_0}$, potential outcomes depend on the forcing variable solely through the treatment indicator, z, but not directly, so that, values of the forcing variable leading to the same treatment status define the same potential outcome. The key implication of Assumption 2 is that it allows us to write $Y_i(\mathbf{s})$ as $Y_i(z_i)$ for each unit $i \in \mathcal{U}_{s_0}$, avoiding to define potential outcomes as functions of the forcing variable. Therefore under local RD-SUTVA for each unit within $\mathcal{U}_{s_0}$ there exist only two potential outcomes, $Y_i(0)$ and $Y_i(1)$: they are the values Y if the unit had a value of the forcing variable falling above and below the threshold, respectively.

Local RD-SUTVA is an important limitation and its plausibility depends on the substantive meaning of the forcing variable and on the support of S for each unit. It may be plausible for the subpopulations $\mathcal{U}_{s_0}$, comprising units who have a relatively large probability that the realized values of S fall in a neighborhood around s_0, but it is arguably plausible for the whole study population, and this may be a major obstacle to the generalization of results from RD designs.

Under local RD-SUTVA, causal effects are defined as comparisons of the potential outcomes $Y_i(0)$ and $Y_i(1)$ for a common set of units in $\mathcal{U}_{s_0}$. They are local causal effects in that they are causal effects for units belonging to a subpopulation $\mathcal{U}_{s_0}$. Typical causal estimands of interest in RD designs are average treatment effects. If focus is on the finite population $\mathcal{U}_{s_0}$, then the average treatment effect is the Sample Average Treatment Effect defined as

$$\tau_{s_0}^S = \frac{1}{N_{\mathcal{U}_{s_0}}} \sum_{i \in \mathcal{U}_{s_0}} [Y_i(1) - Y_i(0)]$$

If the $N_{\mathcal{U}_{s_0}}$ units are considered as a random sample from a large superpopulation (where Assumptions 1 and 2 hold), the causal estimand of interest is the Population Average Treatment Effect:

$$\tau_{s_0} = \mathbb{E}\left[Y_i(1) - Y_i(0) | i \in \mathcal{U}_{s_0}\right]$$

Statistical inference for causal effects requires the specification of an assignment mechanism, i.e., the process that determines which units has a value of the forcing variable falling above or below the threshold, and so which potential outcomes are realized and which are missing. In our approach to RD designs the assignment mechanism is a probabilistic model for the conditional probability of the forcing variable given potential outcomes and covariates. The specification of this assignment mechanism is the distinguishing feature of the approach we propose. Specifically, we formalize the concept of a RD design as local randomized experiment invoking the following assumption:

Assumption 3 *(Local randomization) For each* $i \in \mathcal{U}_{s_0}$*,*

$$Pr\left(S_i | \mathbf{X}_i, Y_i(0), Y_i(1)\right) = Pr(S_i)$$

where $\mathbf{X}_i$ *is a vector of pretreatment variables.*

Note that Assumption 3 can be relaxed assuming that local randomization holds conditional on pretreatment variables, and the analysis of RD designs under ignorable assignment mechanisms given covariates is a valuable topic for future research. This is an assumption similar to those considered in Mealli and Rampichini (2012); Angrist and Rokkanen (2015) and Keele et al. (2015).

Assumption 3 implies that for each unit $i \in \mathcal{U}_{s_0}$, $Pr(S_i \leq s_0 | \mathbf{X}_i, Y_i(0), Y_i(1)) = Pr(S_i \leq s_0) = Pr(Z_i = 1)$, which amounts to state that within the subpopulation $\mathcal{U}_{s_0}$ a Bernoulli trial has been conducted, with individual assignment probabilities depending only on the distribution of the forcing variable, not on either the potential outcomes or pretreatment variables. In other words, Assumption 3 implies that the treatment is randomly assigned in some small neighborhood, $\mathcal{U}_{s_0}$, around s_0, formalizing the key idea by Thistlethwaite and Campbell (1960) that a "true" experiment has been conducted in a neighborhood of the threshold (Thistlethwaite and Campbell, 1960, page 310).

3. Inference on Local Causal Effects for a Subpopulation $\mathcal{U}_{s_0}$

3.1 Selection of subpopulations $\mathcal{U}_{s_0}$

Assumptions 1-3 amount to assuming that within subpopulations $\mathcal{U}_{s_0}$ a classical randomized experiment has been conducted, therefore if at least a true subpopulation $\mathcal{U}_{s_0}$ were known, we could draw inference on causal effects for the subpopulation $\mathcal{U}_{s_0}$ using standard methods for analyzing randomized experiments (e.g., Imbens and Rubin, 2015). Unfortunately, in practice, the true subpopulations $\mathcal{U}_{s_0}$ are usually unknown. Therefore an important issue, in practice, is the selection of a subpopulation $\mathcal{U}_{s_0}$.

In principle, a subpopulation may come in any shape or form. Following Li et al. (2015), we limit our choice to symmetric intervals around s_0 for convenience, assuming that for units belonging to a supposedly existing subpopulation $\mathcal{U}_{s_0}$, the realized value of the forcing variable falls in a symmetric interval around the threshold. Formally, we assume:

Assumption 4 *There exists $h > 0$ such that for each $\epsilon > 0$, $Pr(s_0 - h \leq S_i \leq s_0 + h) > 1 - \epsilon$, for each $i \in \mathcal{U}_{s_0}$.*

Recall that Assumptions 1-3 (and Assumption 4) do not imply that $\mathcal{U}_{s_0}$ has to be unique, therefore we are not interested in finding the largest h, but we only aim at determining plausible values for h.

It is worth noting that the bandwidth choice problem also arises in more conventional RD approaches but for a very different objective. In standard RD approaches, where focus is on estimating causal effects at the threshold, neighborhood selection approaches are usually based on criteria related to local or global polynomial regression methods used to approximate the unknown conditional expectations of the potential outcomes and to obtain an "optimal" extrapolation towards the threshold (see Cattaneo and Vazquez-Bare, 2016, for a review of these methods). In our framework, the objective is to find a subpopulation where Assumptions 1 through 3 are plausible. Consistently the approach for selecting bandwidths h we proposed in Li et al. (2015) exploits Assumption 3. Assumption 3 is a "local" randomization assumption, in the sense that it holds for a subset of units, but may not hold in general for other units. Specifically, Assumption 3 implies that within a subpopulation $\mathcal{U}_{s_0}$ all observed and unobserved pretreatment variables are well balanced in the two subsamples defined by assignment, Z. Therefore, under the assumption that all relevant variables known (or believed) to be related to both treatment assignment and the potential outcomes are observed, within a subpopulation $\mathcal{U}_{s_0}$ any test of the null hypothesis of no effect of assignment on covariates should fail to reject the null. Rejection of the null hypothesis can be interpreted as evidence against the local randomization assumption, at least for the specific subpopulation at the hand. Cattaneo et al. (2015) also exploits balance tests of covariates to select a suitable subpopulation around the threshold, but their approach aims at selecting the largest subpopulation.

Assessing balance in the observed covariates raises problems of multiple comparisons, which may lead to a much higher than planned type I error if they are ignored (e.g., Benjamini and Hochberg, 1995). Cattaneo et al. (2015) prefer to take a conservative approach, by conducting tests for the null hypothesis of balance for each covariate separately, and ignoring the problem of multiplicities. We believe that it may be extremely valuable to account for multiplicities in RD settings, also to avoid to end up with overly small subpopulations.

In the literature, there exist several approaches to tackle the problem of multiple comparisons. From a causal inference perspective, we can use a randomization-based mode of inference, and implement randomization tests adjusted for multiplicities (Lee et al., 2016). As an alternative we can opt for a Bayesian model-based approach, using a Bayesian multiple testing method (e.g., Berry and Berry, 2004; Scott and Berger, 2006). The Bayesian procedure provides a measure of the risk (posterior probability) that a chosen interval around the threshold defines a subpopulation of units that does not exactly matches any true subpopulation, including subjects for which Assumptions 1 through 3 do not hold (see Li et al., 2015).

3.2 Inference

Once subpopulations where Assumptions 1 through 3 are plausible have been selected, we can move to the analysis phase, using any procedure for estimating causal effects from classical randomized experiments, including randomization-based or Bayesian model-based modes of inference.

Randomization inference and Bayesian methods, not relying on asymptotic approximations, are particularly attractive in RD settings where the analysis may rely on a small sample size. Randomization inference provides exact inferences for the finite selected population $\mathcal{U}_{s_0}$, focusing on finite sample causal estimands. From a Bayesian perspective, all inferences are based on the posterior distributions of causal estimands, which are functions of potential outcomes. Therefore inference about sample-average and population-average estimands can be drawn using the same inferential procedures.

A model-based approach requires to specify a model for the potential outcomes. It is worth noting, however, that modeling assumptions play a distinctive role in our setting. They are not necessary and are mainly introduced to adjust for covariates and improve inference: In our setting, model assumptions essentially play the same role as in classical randomized experiments. Conversely, model assumptions are generally crucial in conventional approaches to RD design, where focus is on specifying 'optimal' functional forms relating the outcome to the forcing variable to draw inference on causal effects at the threshold.

Adjusting for both pretreatment variables and the realized values of the forcing variable may be valuable in our approach to RD designs. If the true subpopulations $\mathcal{U}_{s_0}$ were known, in theory, we would not need to adjust for S, because local randomization guarantees that for units in $\mathcal{U}_{s_0}$ values of the forcing variable falling above or below the threshold are independent of the potential outcomes. Nevertheless, in practice, the true subpopulations $\mathcal{U}_{s_0}$ are usually unknown and the risk that a chosen interval around the threshold defines a subpopulation that includes units not belonging to the any true subpopulation, $\mathcal{U}_{s_0}$, is not zero. Systematic differences in the forcing variable S that, by definition, occur between treatment groups may affect inference in the presence of units who do not belong to any subpopulation $\mathcal{U}_{s_0}$. Therefore in order to account for the presence of these units, it might be sensible to conduct inference conditioning on both covariates and the realized values of the forcing variable.

Covariates and forcing variable can be easily incorporated in a Bayesian approach, and they may also help reduce posterior variability of the estimates. Adjusting for S, and possibly for covariates, may be more difficult in randomization-based inference, even if there exist some results in the literature that may be fruitfully exploited in our RD setting (Rosenbaum, 2002; Conti et al., 2014).

4. An Illustrative Example: The Effect of University Grants on Dropout

We illustrate our framework in an example concerning the impact of University student-aid policies on academic careers, using data from the cohort of first-year students enrolled between 2004 to 2006 at University of Pisa and University of Florence (Italy). In Italy, state universities offer grants every year to a limited number of eligible freshmen. In order to get a grant, a student must both meet some eligibility criteria, which are based on an economic indicator of the student's family income and assets falling below or above a

prefixed threshold, as well as apply for the grant. Therefore the grant assignment rule appeals to a RD design, with the economic indicator acting as the forcing variable. Let S be the student's family economic indicator.

In this study, for simplicity, we focus on the effect of eligibility, thus neglecting both the application status and the actual receipt of the grant. The effect of eligibility must be interpreted as an intention-to-treat effect (ITT). The eligibility rule appeals to a *sharp* RD design: Students are eligible if their family economic indicator is below the threshold of 15 000 Euros, and are ineligible otherwise. Therefore for each student i the eligibility indicator is equal to $Z_i = \mathbf{1}\{S_i \leq 15\,000\}$. The outcome variable of primary interest is dropout at the end of the first year. Let $Y_i(z)$ be an indicator for dropout given eligibility status z, and let $Y_i = Y_i(Z_i)$ be the actual dropout indicator observed. In addition, a vector of pretreatment variables, $\mathbf{X}_i$, is observed for each student.

Table 1 presents means for the sample of 16 361 students grouped by eligibility status, Z_i. Eligible freshmen, including students from very low-income families, show different characteristics from ineligible students: on average they have lower high-school grades, and are less likely to come from a science high school and to choose a technical major in University.

We first apply the regression-based approach proposed by Thistlethwaite and Campbell (1960). We divide the forcing variable into evenly-spaced bins and calculate the proportion of students dropping out in each bin. Then, we fit linear regression functions to the observations on either side of the cutoff point, under the assumption that there exists a linear relationship between the outcome (dropout) and the forcing variable.

Figure 1 presents the results. As we can see in Figure 1, there exists a discontinuity at the threshold, which can be interpreted as average treatment effect of eligibility at the threshold according to the original heuristic reasoning of Thistlethwaite and Campbell (1960). The estimate of the ITT effect at the threshold based on the linear regression approach is approximately equal to -0.037%, suggesting that the eligibility reduces dropout for students from families with a value of the economic indicator near the threshold.

Since the publication of Thistlethwaite and Campbell's paper in the early sixties the literature has evolved, and regression or modeling assumptions have been replaced by smoothness/continuity assumptions on the relationship between the outcome and the forcing variable. Table 2 shows estimates of, and 95% confident intervals for, the (population) ITT effects at the threshold derived under the assumption that the conditional distribution functions of the potential outcomes given the forcing variable are continuous in the forcing variable at the threshold. We apply local polynomial estimators, using both a rectangular and a triangular kernel, where the smoothing parameter, the bandwidth, is selected using modern fully data-driven methods, namely, the Coverage Error Rate (CER)-optimal bandwidth proposed by Calonico et al. (2016), used to derive confidence intervals for the average causal effect at the threshold, and two Mean Square Error (MSE)-optimal bandwidths, the Imbens-Kalyanaraman (IK) optimal bandwidth proposed by Imbens and Kalyanaraman (2012) and an upgraded version of it proposed by Calonico et al. (2014). For illustrative purposes, in Table 2 we focus on estimates based on standard local polynomial estimators. Nevertheless, estimates from bias-corrected/robust local polynomial estimators can be also easily applied (see, e.g., Calonico et al., 2014, for details).

Table 1: Italian University Grant Study: Summary Statistics

Variable	All ($n = 16\,361$)	$Z = 0$ ($n = 4\,281$)	$Z = 1$ ($n = 12\,080$)
Assignment variables			
Forcing variable (S)	11148.16	17373.12	8942.12
Grant receipt status (Z)	0.74	0.00	1.00
Outcome variable			
Dropout (Y)	0.38	0.36	0.39
Pre-treatment variables ($\mathbf{X}$)			
Gender	0.60	0.58	0.60
High School Type			
Humanity	0.27	0.26	0.27
Science	0.30	0.36	0.28
Tech	0.39	0.36	0.40
Other	0.05	0.02	0.05
High School grade	81.13	81.94	80.84
Year			
2004	0.40	0.40	0.39
2005	0.34	0.36	0.34
2006	0.26	0.23	0.27
University (Pisa)	0.42	0.39	0.43
Major in University			
Humanity	0.23	0.22	0.23
Social Science	0.26	0.23	0.26
Science	0.13	0.13	0.13
Bio-Med	0.14	0.14	0.14
Tech	0.19	0.22	0.18
Other	0.06	0.06	0.06

Figure 1: Regression of dropout on the forcing variable

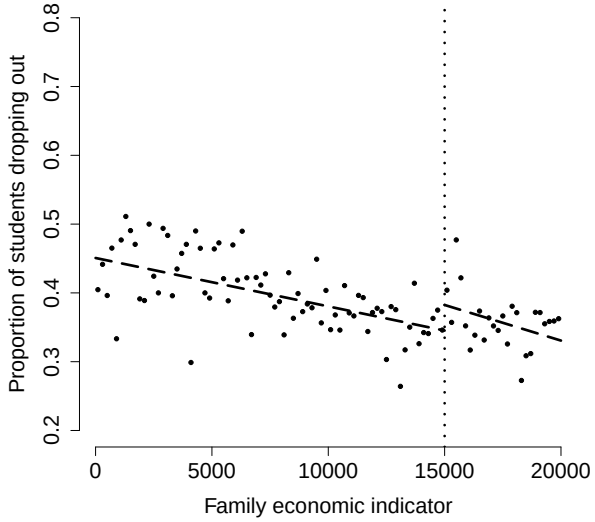

Table 2: Italian University Grant Study: Local Polynomial Estimates of the ITT Effect

Local polynomial	Rectangular Kernel			Triangular Kernel		
regression of order p	τ_{s_0}	*s.e.*	95% CI	τ_{s_0}	*s.e.*	95% CI
CER-optimal bandwidth $= 1\,316.695\,(n = 2\,796)$						
$p = 0$	-0.034	0.018	$[-0.069;\ 0.002]$	-0.045	0.021	$[-0.087; -0.003]$
$p = 1$	-0.066	0.037	$[-0.138;\ 0.006]$	-0.056	0.040	$[-0.134;\ 0.023]$
$p = 2$	-0.039	0.054	$[-0.145;\ 0.067]$	-0.039	0.058	$[-0.152;\ 0.074]$
MSE-optimal bandwidth $= 2\,138.827\,(n = 4\,451)$						
$p = 0$	-0.027	0.014	$[-0.056;\ 0.001]$	-0.032	0.017	$[-0.065;\ 0.001]$
$p = 1$	-0.041	0.029	$[-0.098;\ 0.016]$	-0.057	0.032	$[-0.119;\ 0.005]$
$p = 2$	-0.082	0.043	$[-0.166;\ 0.003]$	-0.068	0.046	$[-0.157;\ 0.022]$
IK optimal bandwidth $= 3\,619.086\,(n = 7\,346)$						
$p = 0$	-0.005	0.011	$[-0.027;\ 0.017]$	-0.022	0.013	$[-0.047;\ 0.004]$
$p = 1$	-0.054	0.022	$[-0.098; -0.011]$	-0.051	0.025	$[-0.099; -0.003]$
$p = 2$	-0.045	0.033	$[-0.111;\ 0.021]$	-0.056	0.036	$[-0.126;\ 0.014]$

As we can see in Table 2, the results are quite sensitive to the choice of both the bandwidth and the polynomial order. The size of the effects changes substantially across different bandwidths, although most of the 95% confidence intervals includes zero. Esti-

mates are also rather unstable across different polynomial orders, especially when the MSE- and IK-optimal bandwidths are used. Nonzero-order polynomials lead to estimate somewhat larger effects than the zero-order polynomial. In some scenario even the choice of the kernel makes a difference. For instance, when the IK-optimal bandwidth and zero-order polynomial are used, the size of the estimate based on the rectangular kernel is about 1/5 of that based on the triangular kernel (-0.005 versus -0.022).

The high sensibility of the inferential results to the critical choices underlying standard RD analyses casts serious doubts on the credibility of the estimates. We argue that these results might strongly rely on extrapolation and model assumptions, especially if the local randomization assumption does not hold for subpopulation of students with a value of the forcing variable falling within a neighborhood defined by some optimal bandwidths, such as the MSE- or IK-optimal bandwidth.

We finally apply the approach we propose, starting by selecting suitable subpopulations $\mathcal{U}_{s_0}$ (see Section 3). We apply randomization-based tests with adjustment for multiplicities to find subpopulations of units where our RD assumptions are plausible. All the covariates listed in Table 1 are considered and we believe that they include all relevant potential confounders.

Table 3 shows randomization-based adjusted p-values for the null hypotheses that the covariates have the same distribution between treated and untreated students for subpopulations defined by various bandwidths, included the optimal bandwidths used in the standard RD analysis. Table 3 also shows $p-$values for the whole sample with S between 0 to 20 000 Euros (column named "ALL") for comparison.

All variables are well balanced for subpopulations defined by bandwidths strictly lower than 1 500. For larger subpopulations some covariates, such as the "indicator of university" are clearly unbalanced. Therefore reasonable subpopulations include students with realized values of the forcing variable within at most 1 500 Euro around the threshold. It is worth noting that only the CER-optimal bandwidth is lower than 1 500 Euro; the MSE- and IK-optimal bandwidths are larger, and define subpopulations where there is clear evidence that covariates are significantly different between eligible and ineligible students. This imbalance justifies, at least partially, the high sensibility of standard RD results to the choice of the bandwidth and model assumptions.

Given the selected subpopulations $\mathcal{U}_{s_0}$, we use a Neyman approach for inference. Table 4 shows estimates of, and 95% confidence intervals (based on the Normal approximation) for the ITT effect for bandwidths ranging from 500 to 1 500 Euros. The estimated ITT effects are similar across different bandwidths: All the estimates are negative, suggesting that eligibility reduces dropout, but most of them are not significant at the 5% level. Only for the subpopulation of students within 1 000 Euros around the threshold, the 95% confidence interval do not cover zero. For this subpopulation the estimated ITT effect of eligibility is a reduction in dropout rate of about 4.7%. The precision of the estimates could be improved adjusting for covariates using a model-based Bayesian approach, which involves model assumptions. Recall that, however, under our framework, identification does not rely on model assumptions; they are only used to improve inference.

Table 3: Italian University Grant Study: Adjusted p-values for the null hypothesis that covariates have the same distribution between eligible and ineligible students for various subpopulations

	Local Randomization Bandwidths						*Local Polynomial Bandwidths*		
Variable	500	1000	1500	2000	5000	All	1316.695	2138.827	3619.086
(Sample size)	(1 042)	(2 108)	(3 166)	(4 197)	(9 846)	(16 361)	(2 796)	(4 451)	(7 346)
Gender	1.000	1.000	1.000	1.000	0.307	0.058	1.000	1.000	0.953
High School Type									
Humanity	1.000	0.999	1.000	1.000	0.973	0.996	1.000	1.000	1.000
Science	1.000	1.000	1.000	0.909	0.001	0.001	1.000	0.686	0.227
Tech	1.000	1.000	1.000	1.000	0.084	0.001	0.998	1.000	1.000
Other	0.432	0.720	0.402	0.281	0.004	0.001	0.541	0.250	0.081
High School Grade	0.991	1.000	1.000	1.000	1.000	0.001	1.000	1.000	1.000
Year									
2004	1.000	1.000	1.000	1.000	1.000	0.987	1.000	1.000	1.000
2005	1.000	0.943	1.000	1.000	0.847	0.066	0.999	1.000	0.877
2006	1.000	1.000	1.000	1.000	0.788	0.001	1.000	1.000	0.939
University (Pisa)	0.998	1.000	0.117	0.006	0.001	0.001	0.602	0.018	0.004
Major in University									
Humanity	0.965	0.295	0.405	0.910	0.969	0.970	0.562	0.955	1.000
Science	1.000	1.000	0.999	0.998	1.000	1.000	1.000	0.995	0.991
Social Science	1.000	1.000	1.000	1.000	0.998	0.001	1.000	1.000	1.000
Bio-Med	0.995	0.698	0.999	0.999	1.000	1.000	0.992	0.990	1.000
Tech	0.965	0.984	1.000	1.000	0.123	0.001	1.000	1.000	0.717
Other	0.989	1.000	1.000	1.000	0.858	0.993	1.000	1.000	1.000

Table 4: Italian University Grant Study: Estimates of, and 95% Confidence Intervals for, the ITT Effect for various subpopulations $\mathcal{U}_{s_0}$ based on Neyman's approach

Bandwidth	τ^S_{s0}	*s.e.*	95% CI	
500	−0.026	0.030	[−0.085;	0.033]
1000	−0.047	0.021	[−0.088;	−0.005]
1500	−0.020	0.017	[−0.054;	0.014]
CER-optimal bandwidth				
1316.695	−0.034	0.018	[−0.069;	0.002]

5. Discussion

There exist alternative approaches to formalize and analyze RD designs as local randomized experiments. Simultaneously with, but separately from Li et al. (2015), Cattaneo et al. (2015) and Sales and Hansen (2015) propose different sets of assumptions within a neighborhood of the threshold that allow one to exploit a local randomization assumption as an identification and estimation strategy in RD designs. Our approach presents subtle but important differences with the methodological framework proposed by Cattaneo et al. (2015) and Sales and Hansen (2015). In particular, we develop a framework for RD analysis that is fully consistent with the potential outcome approach, by clearly defining the treatments and potential outcomes and separating and defining the critical assumptions.

Sales and Hansen (2015) propose to use regression methods to remove the dependence of the outcome from the forcing variable, and then assume that the transformed version of the outcome is independent of treatment assignment, that is, Z in our notation.

The key assumption in Cattaneo et al. (2015) – named 'local Randomization' – does not actually define an assignment mechanism as the conditional probability of the assignment variable given covariates and potential outcomes, which is the general definition of assignment mechanism in the potential outcome approach to causal inference (Imbens and Rubin, 2015). The local randomization assumption proposed by Cattaneo et al. (2015) has two components. The first component amounts to assuming that the marginal distributions of the forcing variable are the same for all units inside a specific subpopulation. This assumption does not formally define an assignment mechanism but simply implies that the values of the forcing variable can be considered "as good as randomly assigned." The second component requires that potential outcomes depend on the values of the forcing variable only through treatment indicators. We view this assumption as part of SUTVA, that is, as part of the definition of potential outcomes, rather than as an assumption on the assignment mechanism.

The birth of these alternative interpretations and formalizations of a RD designs has raised some discussion on the relationship between local randomization and continuity RD assumptions (e.g., de la Cuesta and Imai, 2016; Skovron and Titiunik, 2015).

It is worth noting that in approaches to RD designs where the forcing variable is viewed as a pre-treatment covariate, the conditional independence assumption trivially holds, but it cannot be exploited directly due to the violation of the overlap assumption. In these settings some kind of extrapolation is required, and in order to avoid that estimates heavily rely on extrapolation, previous analyses focus on causal effects of the treatment for units at the threshold under smoothness assumptions, such as continuity assumptions.

Some authors (de la Cuesta and Imai, 2016; Cattaneo et al., 2015; Skovron and Titiunik, 2015; Sekhon and Titiunik, 2016) argue that the local randomization assumption is not required for the RD design. According to us, this sentence may be misleading and deserves some discussion.

Continuity assumptions and our local randomization assumption are different assumptions: they lead to identify and estimate different causal estimands. Local randomization is not required to identify and estimate causal effects *at the threshold*, the causal effects typically considered in the RD design literature, but it is required to identify and estimate causal effects *around the threshold*.

Although causal effects at the threshold are identified under continuity assumptions, which imply that randomization took place precisely at the threshold, we argue that inference under local randomization may be more robust. Specifically, even if focus is on causal effects at the threshold, and continuity assumptions are invoked for inference, in practice, in any analysis of data we are always forced to actually use information on units that are far away from the threshold, relying on some form of extrapolation. In the literature, the choice of a sub-sample of units in a neighborhood of the threshold is usually based on local or global polynomial regression approximations of the unknown conditional expectations of the potential outcomes given the forcing variable. Recently fully data-driven methods, based on selecting an optimal bandwidth under squared error loss (for the local-linear regression estimator, the local polynomial estimator and generalizations) have become increasingly popular (Imbens and Kalyanaraman, 2012; Calonico et al., 2014). These methods do not guarantee, however, that units with a value of the forcing variable falling above and below the threshold have similar distributions of the covariates. If pre-treatment variables are not well-balanced between units above and below the threshold, inference may be highly sensitive to functional assumptions, such as the choice of a local estimator, that is, the choice of the weights from the kernel defining the local estimator. Conversely, if the local randomization assumption holds, and the neighborhood around the threshold is selected aiming at choosing a sub-sample of units where pre-treatment variables are well-balanced between units above and below the threshold, we expect that inference is robust with respect to model assumptions, including the choice of kernels of local estimators. This is analogous to the result about consistency of regression-based estimates of average causal effects from randomized experiments, where consistency does not rely on the linearity of the relationship between outcome, treatment and covariates (Imbens and Rubin, 2015, Chapter 7).

Under local randomization causal estimands of interest are causal effects for units belonging to a sub-population $\mathcal{U}_{s_0}$, which generally includes units with values of the forcing variable falling in a neighborhood "away" from the threshold. Therefore, under local randomization we can identify and estimate causal effects away from the threshold. Alternative ways to generalize RD results aways from the cutoff point require additional ignorability-type assumptions (e.g., Battistin and Rettore, 2008; Mealli and Rampichini, 2012; Angrist and Rokkanen, 2015). Mealli and Rampichini (2012) combine unconfoundedness and differences-in-differences approaches to extend estimates of causal effects from RD analyses away from the cutoff point. Ways to further exploiting randomization-type assumptions to generalize results from RD analyses away from the threshold are still under investigation.

Acknowledgments

The authors acknowledge financial support from the Italian Ministry of Research and Higher Education through grant Futuro in Ricerca 2012 RBFR12SHVV_003.

References

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

Athey, S. and Imbens, G. (2016). The state of applied econometrics - causality and policy evaluation. *ArXiv working paper*, No 1607.00699.

Battistin, E. and Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, 142:715–730.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society - Series B*, 57:289–300.

Berry, S. M. and Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60:418–426.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016). On the effect of bias estimation on coverage accuracy in nonparametric inference. *arXiv Working paper: 1508.02973*, 82.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Campbell, D. (1969). Reforms as experiments. *American Psychologist*, 24:409–442.

Cattaneo, M. and Escanciano, J. C. (2016). Regression discontinuity designs: Theory and applications. *Advances in Econometrics*, 38. Emerald Group Publishing. To appear.

Cattaneo, M., Frandsen, B. R., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. and Vazquez-Bare, G. (2016). The choice of neighborhood in regression discontinuity designs. *Observational Studies*, 2:134–146.

Constantinou, P. and O'Keeffe, A. G. (2016). Regression discontinuity designs: A decision theoretic approach. *ArXiv working paper*, No 1601.00439.

Conti, E., Duranti, S., Mattei, A., Mealli, F., and Sciclone, N. (2014). The effects of a dropout prevention program on secondary students' outcomes. *RIV Rassegna Italiana di Valutazione*, 58:15–49.

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

Dawid, A. P. (2000). Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Society*, 95:407–448.

de la Cuesta, B. and Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19:375–396.

Dinardo, J. and Lee, D. S. (2011). Program evaluation and research designs. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4A, pages 463–536. Elsevier Science B.V.

Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3).

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

Imbens, W. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction.* Cambridge University Press, New York, NY, USA.

Keele, L., Titiunik, R., and Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical. Society, Series A*, 178(1).

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 485:281–355.

Lee, J. J., Miratrix, L., Pillai, N. S., and Forastiere, L. (2016). More powerful multiple testing in randomized experiments with non-compliance. *Statistica Sinica*, To appear.

Li, F., Mattei, A., and Mealli, F. (2015). Bayesian inference for regression discontinuity designs with application to the evaluation of italian university grants. *The Annals of Applied Statistics*, 9(4):1906–1931.

Mealli, F. and Rampichini, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *Journal of the Royal Statistical. Society, Series A*, 175(3):775–798.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–304.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in the fisher randomization test" by basu. *Journal of the American Statistical Association*, 75:591–593.

Sales, A. and Hansen, B. B. (2015). Limitless regression discontinuity: Causal inference for a population surrounding a threshold. *ArXiv working paper*, No 1403.5478.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136:2144–2162.

Sekhon, J. S. and Titiunik, R. (2016). On interpreting the regression discontinuity design as a local experiment. *Advances in Econometrics*, 38. Emerald Group Publishing. To appear.

Skovron, C. and Titiunik, R. (2015). A practical guide to regression discontinuity designs in political science. *Working paper. Department of Political Science University of Michigan.*

Thistlethwaite, D. and Campbell, D. (1960). Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.

Observational Studies 2 (2016) 174-182 Submitted 10/16; Published 12/16

Understanding Regression Discontinuity Designs As Observational Studies

Jasjeet S. Sekhon **sekhon@berkeley.edu**
Robson Professor
Departments of Political Science and Statistics
UC-Berkeley
210 Barrows Hall #1950, Berkeley, CA 94720-1950

Rocío Titiunik **titiunik@umich.edu**
James Orin Murfin Associate Professor
Department of Political Science
University of Michigan
505 South State St., 5700 Haven Hall,Ann Arbor, MI 48109-1045

Keywords: Regression Discontinuity, Local Randomization, Local Experiment

1. Introduction

Thistlethwaite and Campbell (1960) proposed to use a "regression-discontinuity analysis" in settings where exposure to a treatment or intervention is determined by an observable score and a fixed cutoff. The type of setting they described, now widely known as the regression discontinuity (RD) design, is one where units receive a score, and a binary treatment is assigned according to a very specific rule. In the simplest case, all units whose score is above a known cutoff are assigned to the treatment condition, and all units whose score is below the cutoff are assigned to the control (i.e., absence of treatment) condition. Thistlethwaite and Campbell insightfully noted that, under appropriate assumptions, the discontinuity in the probability of treatment status induced by such an assignment rule could be leveraged to learn about the effect of the treatment at the cutoff. Their seminal contribution led to what is now one of the most rigorous non-experimental research designs across the social and biomedical sciences. See Cook (2008), Imbens and Lemieux (2008) and Lee and Lemieux (2010) for reviews, and the recent volume edited by Cattaneo and Escanciano (2017) for recent specific applications and methodological developments.

A common and intuitive interpretation of RD designs is that the discontinuous treatment assignment rule induces variation in treatment status that is "as good as" randomized near the cutoff, because treated and control units are expected to be approximately comparable in a small neighborhood around the cutoff (Lee, 2008; Lee and Lemieux, 2010). This local randomization interpretation has been extremely influential, and many consider RD designs to be almost as credible as experiments. Although the formal analogy between RD designs and experiments was discussed recently by Lee (2008), the idea that the RD design behaves like an experiment was originally introduced by Thistlethwaite and Campbell, who called a hypothetical experiment where the treatment is randomly assigned near the

cutoff an "experiment for which the regression-discontinuity analysis may be regarded as a substitute" (Thistlethwaite and Campbell, 1960, p. 310). Building on this analogy, Lee (2008) formalized the idea in a continuity-based framework; in addition, Cattaneo et al. (2015) formalized this idea in a Fisherian finite-sample framework. See Cattaneo et al. (2017) and Sekhon and Titiunik (2017) for recent discussions on the connections between both frameworks.

The analogy between RD designs and experiments has been useful in communicating the superior credibility of RD relative to other observational designs, and has focused attention on the need to perform falsification tests akin to those usually used in true experiments. All these developments have contributed to the RD design's rigor and popularity. Despite these benefits, we believe the analogy between RD designs and experiments is imperfect, and we offer a more cautious interpretation in which the credibility of RD designs ranks decidedly below that of actual experiments.

In our view, RD designs are best conceived as non-experimental designs or *observational studies*—i.e., studies where the goal is to learn about the causal effects of a treatment, but the similarity or comparability of subjects receiving different treatments cannot be ensured by construction. Interpreting RD designs as observational studies implies that their credibility must necessarily rank below that of experiments. This, however, does not mean that RD designs are without special merit. Among observational studies, RD designs are one of the most credible alternatives because important features of the treatment assignment mechanism are known and empirically testable under reasonable assumptions.

We justify our view by focusing on three main issues. First, we consider the RD treatment assignment rule, and show that it contains considerably less information than the analogous rule in an experimental assignment. Second, we consider the special role of the score or running variable, in particular the possibility that the score may affect the outcome via post-treatment channels and violate an exclusion restriction that holds by construction in experiments. Finally, we highlight that in order to obtain meaningful conclusions from testing the "empirical implications" of a valid RD design, further assumptions must be made about the data generating process. All these issues support our view that RD designs are observational studies. We do not mean these arguments as a critique of RD designs. Our point is simply that a compelling observational study faces hurdles that are absent in experimental designs, and therefore the analysis and interpretation of RD designs should be done with the same caution as in any other observational study.

2. The RD Treatment Assignment Rule

The fundamental feature of RD designs is that the treatment is assigned based on a known rule. In the so-called sharp RD design where compliance with treatment is perfect, treatment status is deterministic given the score: all units with score below the cutoff are assigned to and receive the control condition, and all units with score above the cutoff are assigned to and receive the treatment condition. Moreover, in the standard RD setup, the cutoff is known. This can be formalized in the rule $T_i = \mathbb{1}\{X_i \geq c\}$, where $i = 1, 2, \ldots n$ indexes the units in the study, T_i is the treatment status, c is the cutoff, and X_i is the score or running

variable. Because this rule is at the heart of every RD design,[1] any researcher working with an RD design has rich information about the treatment assignment mechanism.

At first glance, the fact that treatment assignment is based on a known rule might suggest that RD designs are not observational studies. As commonly defined (e.g. Rosenbaum, 2002), a key feature of an observational study is that the treatment assignment mechanism is not under the control of the researcher (or someone else the researcher has access to), which implies that it is fundamentally unknown. For example, an observational study of the effects of smoking on lung cancer may compare smokers and non-smokers and obtain valid inferences under some assumptions, but the probability of smoking always remains unknown.

RD designs are different in this regard because, although the actual assignment of treatment is rarely under the direct control of the investigator, the probability of receiving treatment given the score is known for every unit. In other words, if a unit receives a particular score value, in a sharp RD design we know with certainty whether the probability of receiving treatment was one or zero. Although this has many advantages, it is not enough to lift the status of RD from observational studies to experimental designs. The reason is that the distribution of the score remains fundamentally unknown: although we know that $T_i = 1$ if the score X_i is above the cutoff and $T_i = 0$ otherwise, we know nothing about how the value of X_i was determined. Thus, despite the treatment assignment rule being known, the comparability of treated and subjects is not ensured.

This fundamental lack of knowledge about the distribution of the score makes the RD design inherently different from experiments. In an experiment, units are randomly assigned to treatment or control, which implies that the distribution of all predetermined characteristics and unobserved confounders is identical in the treatment and control groups, ensuring their comparability. In the language of the potential outcomes framework, random assignment of treatment ensures independence between treatment status and potential outcomes. In the absence of complications (such as interference across units and compliance issues), this independence is sufficient to guarantee identification of the (sample) average treatment effect.

In contrast, in RD designs, the treatment assignment rule $T_i = \mathbb{1}\{X_i \geq c\}$ is not enough to ensure the identification of the treatment effect (at the cutoff). This is a direct consequence of the fact that the assignment rule determines T_i, but it does not determine X_i. For example, as shown by Hahn et al. (2001), the main condition to obtain identification of the average treatment effect at the cutoff in a sharp RD design is the continuity of the regression functions of the potential outcomes at the cutoff. Letting Y_{1i} and Y_{0i} denote the potential outcomes under treatment and control for unit i, defining the observed outcome as $Y_i = T_i Y_{1i} + (1 - T_i)Y_{0i}$, and assuming the observed data $\{Y_i, X_i\}_{i=1}^n$ is a random sample from a larger population, the continuity condition says that $\mathbb{E}[Y_{1i}|X = x]$ and $\mathbb{E}[Y_{0i}|X_i = x]$, seen as functions of x, are continuous in x at c.

Crucially, the continuity of the potential-outcome regression functions at the cutoff is not implied or guaranteed by the known and deterministic RD treatment assignment rule; it is an assumption that must be imposed. In other words, the fact that the treatment is assigned according to the rule $T_i = \mathbb{1}\{X_i \geq c\}$ places no restrictions on the proper-

1. In a fuzzy RD design compliance with the assignment is no longer perfect; in this case, the rule $T_i = \mathbb{1}\{X_i \geq c\}$ still applies, but T_i now refers to treatment assignment instead of treatment status.

ties of functions such as $\mathbb{E}[Y_{1i}|X = x]$ and $\mathbb{E}[Y_{1i}|X_i = x]$. In contrast, the unconfounded random treatment assignment rule in classical experiments guarantees a statistical independence assumption (or a known randomization distribution assumption in finite-sample settings). This fundamental asymmetry between the credibility of identification conditions in experiments versus RD designs—in the former guaranteed by construction, in the latter by assumption—is one of the reasons why the RD should be considered an observational design.

Randomized experiments do need additional assumptions for parameter estimation and hypothesis testing in many cases. Depending on the parameter or hypothesis of interest and the statistic used, researchers usually need to impose additional regularity conditions, in addition to modeling the sampling structure of the data. For example, in the case of the average treatment effect, these regularity conditions, aside from non-interference, include moment conditions on the outcomes (and covariates)—see, e.g., Lin (2013). Such conditions will typically be weaker than the assumptions required for estimation in the continuity-based RD case, where smoothness conditions are required in addition to the continuity assumption (Calonico et al., 2014), neither of which is guaranteed by the design. We also note that in the case of randomized experiments, both parameter estimation standard hypothesis testing can be skipped in favor of permutation tests of the Fisherian sharp null, which require even weaker assumptions (Rosenbaum, 2002).

3. The Intermediate Role of the Running Variable

The existence of the running variable—and our fundamental lack of knowledge about its distribution and determinants—poses another challenge for the analogy between experiments and RD designs, and gives another reason to classify the latter as an observational design. In a nutshell, the source of this second challenge is that the RD running variable is often a very important determinant of the potential outcomes—not only because it may correlate with predetermined characteristics that are related to the outcome, but also because it can have a "direct" or "post-treatment" effect on the potential outcomes. As we discuss in detail in Sekhon and Titiunik (2017), the special status of the RD score breaks the usual connection between the concepts of random assignment, statistical independence, and constant or "flat" regression functions that are taken for granted in experiments. This exclusion restriction was first noted by Cattaneo et al. (2015) in a Fisherian framework, and is relaxed under additional assumptions in Cattaneo et al. (2017).

One intuitive way to motivate the RD-experiment analogy is that a randomized experiment can be understood as particular case of the RD design where the score is a (pseudo) random number, and the cutoff is chosen to ensure the desired probability of treatment assignment. For example, one can randomly assign a treatment among a group of subjects with probability 50% by assigning a uniform random number between 1 and 100 to each subject, and then assigning the treatment only to those subjects whose assigned number exceeds 50. This randomized experiment can be easily recast as a sharp RD design where the uniform random number is the score and the cutoff is 50.

This hypothetical experiment recast as an RD design has two crucial features:

(i) By virtue of random assignment, the score is statistically independent of all predetermined covariates, including all those covariates that affect or are related to the potential outcomes;

(ii) By virtue of the score being an arbitrary number generated solely for the purpose of assigning the treatment, there can be no "post-treatment" effect of the score on the potential outcomes except via the treatment assignment indicator.

The combination of (i) and (ii) implies, for example, that the regression functions $\mathbb{E}[Y_{0i}|X = x]$ and $\mathbb{E}[Y_{1i}|X_i = x]$ are constant in the entire support of the score.

The RD design, in practice, does not generally satisfy either of these conditions. In typical RD treatment assignment rules, the score or running variable is a crucial determinant of the potential outcomes. For example, a party may win an election when its vote share exceeds 50%, and we may be interested in the effect of winning on future victories. Or a municipality may receive federal assistance when its poverty index is below a certain threshold, and we may be interested in the effect of federal assistance on mortality. In such cases, the score is fundamentally related to both predetermined characteristics of the units that may be strongly related to the outcome (e.g., municipalities with high poverty index may also have high unemployment which can affect mortality via lower health insurance coverage), and it can also affect the outcome directly (e.g., increased poverty may reduce access to potable water and increase disease and mortality risk). Both possibilities make the analogy between experiments and RD designs imperfect.

This challenge can be further illustrated by noting that even if we assume that the score is randomly assigned among subjects, the score—and, consequently, the treatment assignment, may fail to be independent of the potential outcomes. The reason is simply that, although the random assignment of the score ensures condition (i), it fails to ensure condition (ii). A randomly assigned score is by construction independent of all predetermined covariates, but it nonetheless may have an effect on the outcome that occurs not via correlation with predetermined characteristics, but via a post-treatment channel. This implies that the random assignment of the score is not enough to guarantee the exclusion restriction that the score affects the potential outcomes only through the treatment assignment indicator.

To understand why this occurs, note that in a true experiment the exclusion restriction holds by construction because the pseudo-random number assigned to each subject plays no role in the data generating process of the potential outcomes. Importantly, the exclusion restriction holds in a true experiment not because of the random assignment per se, but because the score used to implement the randomization procedure is arbitrary (indeed, in most real experiments, this "score" is entirely unknown to the experimental subjects). This is why in a RD design, where the score may often affect the outcome by various post-treatment channels, the random assignment of the score does not—and cannot—guarantee condition (ii).

This brief discussion shows that assuming random assignment of the RD score in a neighborhood near the cutoff does not imply that the potential outcomes and the treatment are statistically independent, or that the potential outcomes are unrelated to the score in this neighborhood. Furthermore, as we show formally in Sekhon and Titiunik (2017), the assumption of local independence between the potential outcomes and the treatment

assignment does not imply the exclusion restriction that the score affects the outcome only via the treatment indicator but not directly.

In sum, the RD treatment assignment rule does not by itself place any restrictions on the ways in which the score can influence the potential outcomes—and even in a locally random RD design where the score is randomly assigned near the cutoff, the statistical independence between potential outcomes and treatment assignment that we take for granted in experiments need not follow. This is another reason why we view RD designs as observational studies.

4. The RD Assumptions and Their Empirical Implications

Lee (2008) heuristically argued that a consequence of interpreting RD designs as local experiments is that predetermined covariates in treated and control groups should be similar in a neighborhood of the cutoff. Formally, Lee established continuity of the distribution of observed predetermined covariates at the cutoff. As a consequence, he proposed to test whether the treatment has an effect on predetermined covariates at the cutoff to falsify the RD assumptions—similarly to the way in which balance tests are used in experiments to evaluate whether the randomization was performed correctly. This emphasis on the need to test empirically the comparability of treatment and control groups has been a positive and influential development in the RD literature. By now, falsification tests are a standard part of most empirical RD applications (see, e.g., Caughey and Sekhon, 2011; de la Cuesta and Imai, 2016; Eggers et al., 2015).

Under the assumption of continuity of the potential-outcome regression functions, these "covariate balance" tests should be implemented treating each covariate as an outcome in the RD analysis—that is, estimating average RD treatment effects on the covariates in the same way as these effects are estimated for the true outcome of interest. The standard implementation of continuity-based RD estimation and inference uses local polynomial methods, fitting a weighted polynomial of the outcome/covariate on the score within an optimally chosen bandwidth around the cutoff (see, e.g., Calonico et al., 2014, 2016, and references therein). This implementation allows all predetermined covariates to be arbitrarily related to the score variable, and looks for an effect at the cutoff. Since the covariates are determined before treatment is assigned, researchers are reassured when such RD effects on the covariates cannot be distinguished from zero.

The use of these "covariate balance" tests for falsification is perhaps the most salient practical similarity between RD analysis and experimental analysis. The assumption behind the RD falsification tests on covariates is that continuity of the covariate regression functions implies or at least supports the assumption that the potential-outcome regression functions are continuous. This is a strong requirement because, as with continuity of the potential-outcome regression functions, continuity of the covariate regression functions is not implied by the RD treatment assignment rule. Moreover, continuity of the covariate regression functions is neither necessary nor sufficient for the potential-outcome regression functions to be continuous. Thus, falsification tests based on covariates require assumptions that are not true by construction. Similarly, falsification tests based on the density of the running variable (McCrary, 2008) require that such density be continuous at the cutoff, another

condition that is neither necessary nor sufficient for the main RD identification assumption to hold.

It follows that falsification analysis in RD designs is more demanding than in experimental settings. In the case of actual experiments, we know that if the random assignment of the treatment was implemented without errors, the treatment assignment will be independent of all predetermined covariates (as well as of potential outcomes). Thus, the design itself implies that the distribution of predetermined covariates in treatment and control groups is the same, and falsification tests try to corroborate the empirical implication of a balance condition we know to be true. In contrast, in RD designs, neither the identification assumptions on the potential outcomes nor the falsification assumptions on the covariates are known to be true, because these assumptions are not implied by the treatment assignment rule. Precisely for this reason, falsification analysis plays a more crucial role in RD designs than in experiments, as researchers are eager to provide empirical evidence that the invoked RD assumptions are plausible. The paradox is that falsification tests are most needed in those settings where they require more assumptions to be informative. The bottom line is that identification assumptions are a prerequisite for the data to be informative about the parameters of interest, and we cannot use the data to test the assumptions that make the data meaningful in the first place. In general, nonparametric identification assumptions are fundamentally untestable.

This, of course, does not mean that RD falsification tests are not useful. In most applications, it is entirely reasonable to assume that if the potential-outcome regression functions are continuous at the cutoff, most predetermined covariates that are related to the outcome will also have continuous regression functions. This assumption will be particularly plausible for certain covariates, such as the outcome measured before treatment assignment and other variables that are known to be strongly related to the outcome of interest. Our point is simply that this is an assumption that must be made, in contrast to a feature that is true by design.

5. Conclusion

In sum, we believe the RD design is an observational study, and should be interpreted as such. Despite the usefulness of the analogy between RD designs and experiments, RD designs lack the credibility of experiments for the simple reason that the treatment assignment rule does not guarantee the assumptions that are needed for identification of the treatment effects of interest. In particular, the RD assignment rule implies neither continuity of the relevant potential-outcome functions nor local independence between the potential outcomes and the treatment assignment; and the random assignment of the score near the cutoff does not imply local independence between the potential outcomes and the score or treatment assignment. Moreover, falsification tests in RD designs require additional assumptions about the relationship between the selected predetermined covariates and the potential outcomes.

Acknowledgments

We are indebted to Matias Cattaneo, Kosuke Imai, Max Farrell, Joshua Kalla, Kellie Ottoboni, and Fredrik Sävje for valuable comments and discussion. Sekhon gratefully acknowledges support from the Office of Naval Research (N00014-15-1-2367) and Titiunik gratefully acknowledges financial support from the National Science Foundation (SES 1357561).

References

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2016). Regression discontinuity designs using covariates. Working paper, University of Michigan.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cattaneo, M. D. and Escanciano, J. C. (2017). *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing, forthcoming.

Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management,* forthcoming.

Caughey, D. and Sekhon, J. S. (2011). Elections and the regression discontinuity design: Lessons from close U.S. house races, 1942–2008. *Political Analysis*, 19(4):385–408.

Cook, T. D. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654.

de la Cuesta, B. and Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19:375–396.

Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., and Snyder, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, 59(1):259–274.

Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Imbens, G. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York, NY, second edition.

Sekhon, J. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In Cattaneo, M. D. and Escanciano, J. C., editors, Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38). Emerald Group Publishing, forthcoming.

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.

Observational Studies 2 (2016) 183-209 Submitted 10/16; Published 12/16

The Regression Discontinuity Design and the Social Corruption of Quantitative Indicators

Vivian C. Wong **vcw2n@virginia.edu**
Curry School of Education
University of Virginia
Charlottesville, VA

Coady Wing **cwing@indiana.edu**
School of Public and Environmental Affairs
Indiana University
Bloomington, IN

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 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 birthdate. It seems unlikely that many parents would attempt to manipulate measured birthdates 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*.

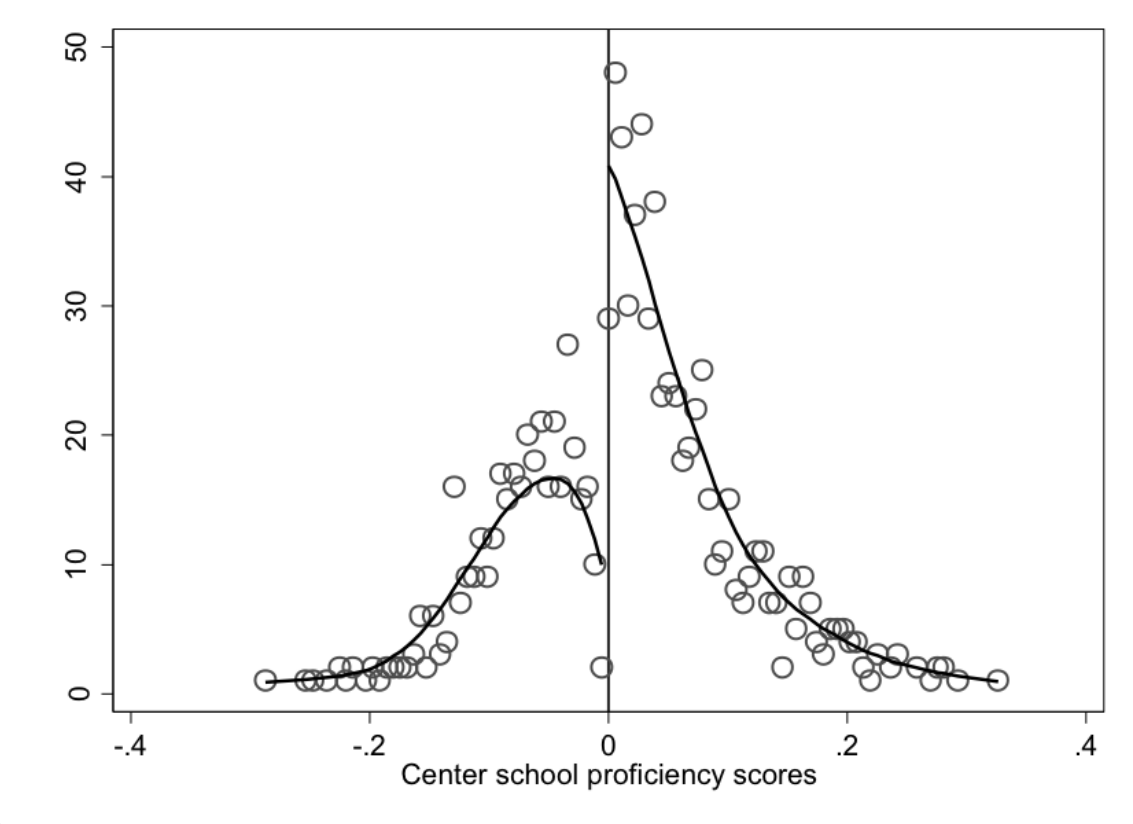

Figure 1: Smoothed frequency of assignment scores in Pennsylvania schools (SWD subgroup only)

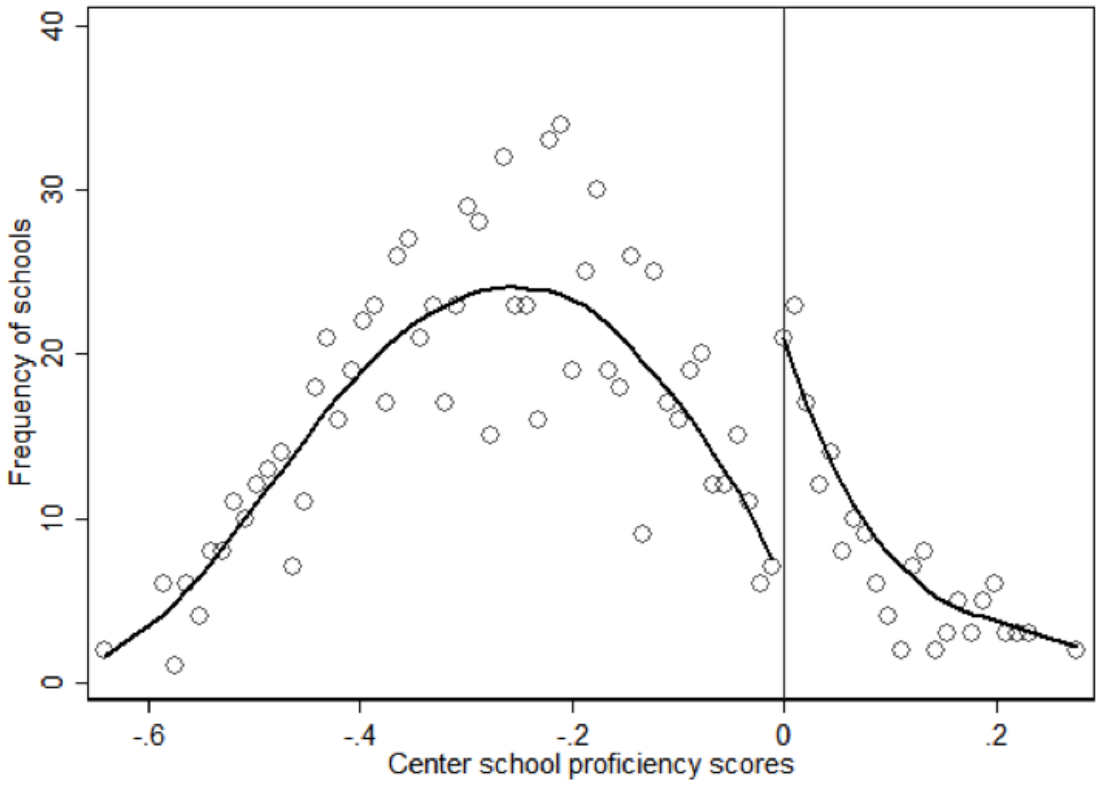

Figure 2: Smoothed frequency of assignment scores with the confidence interval rule as cutoff.

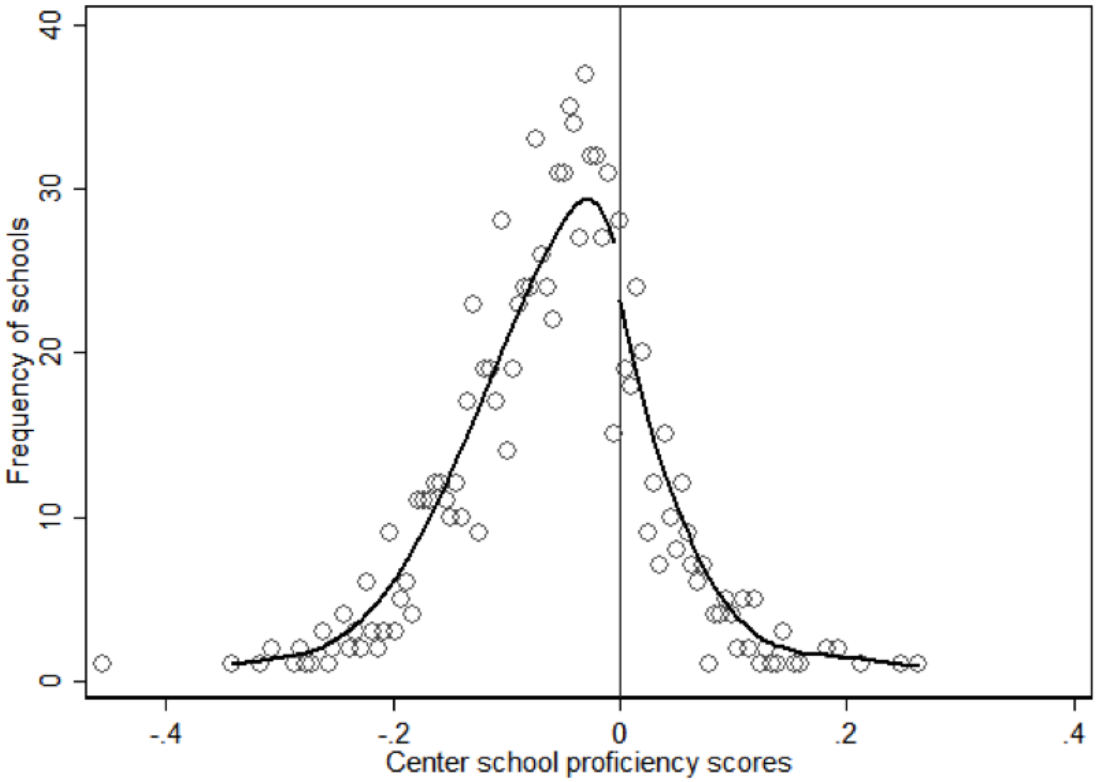

Figure 3: Smoothed frequency of assignment scores with the safe harbor rule as cutoff.

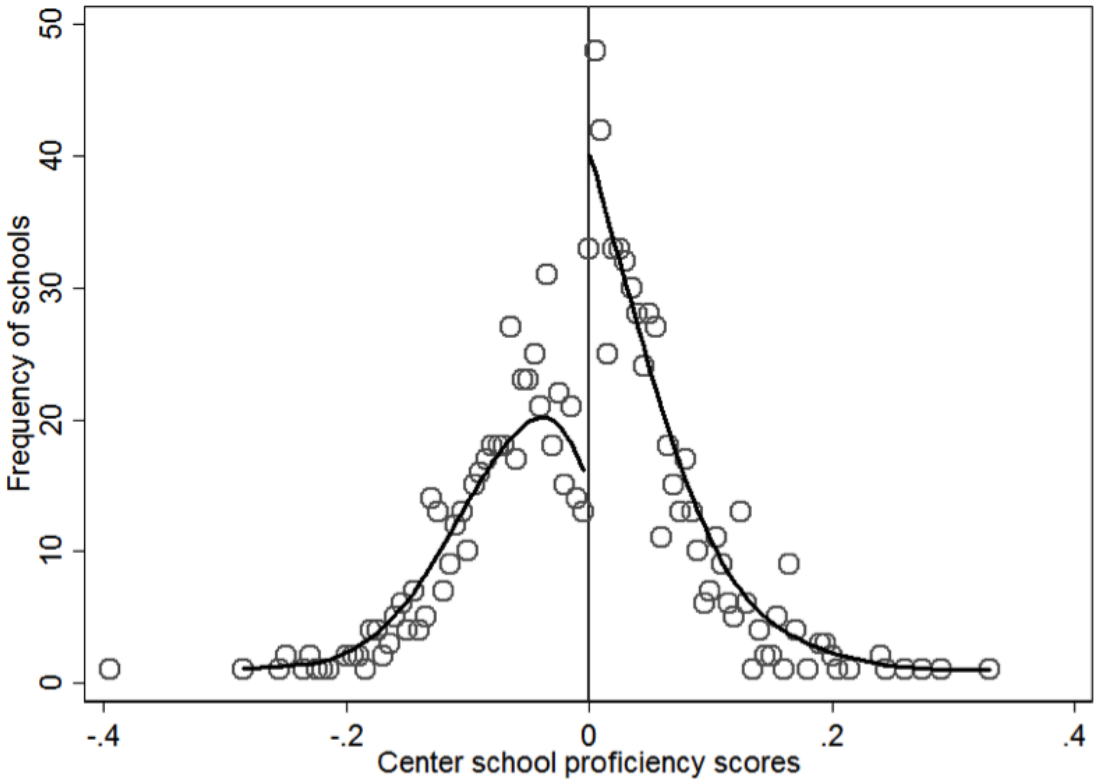

Figure 4: Smoothed frequency of assignment scores with the safe harbor confidence interval rule as cutoff.

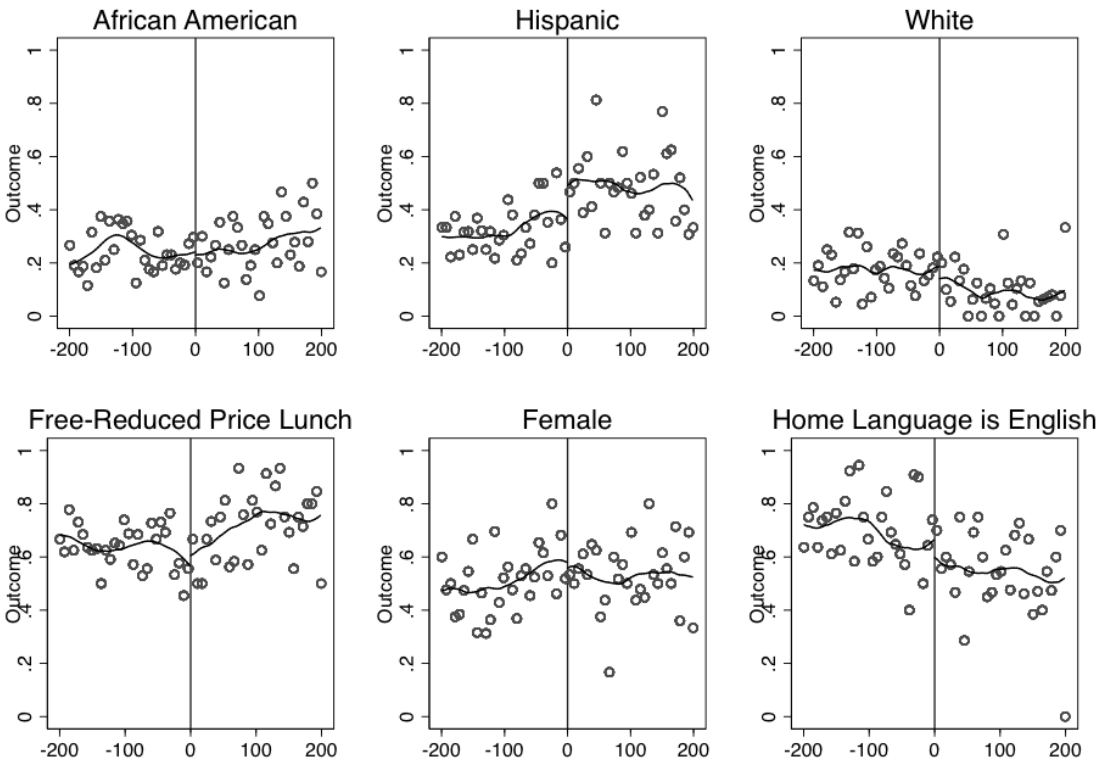

Figure 5: Covariate balance at the cut-off in the New Jersey pre-K sample.

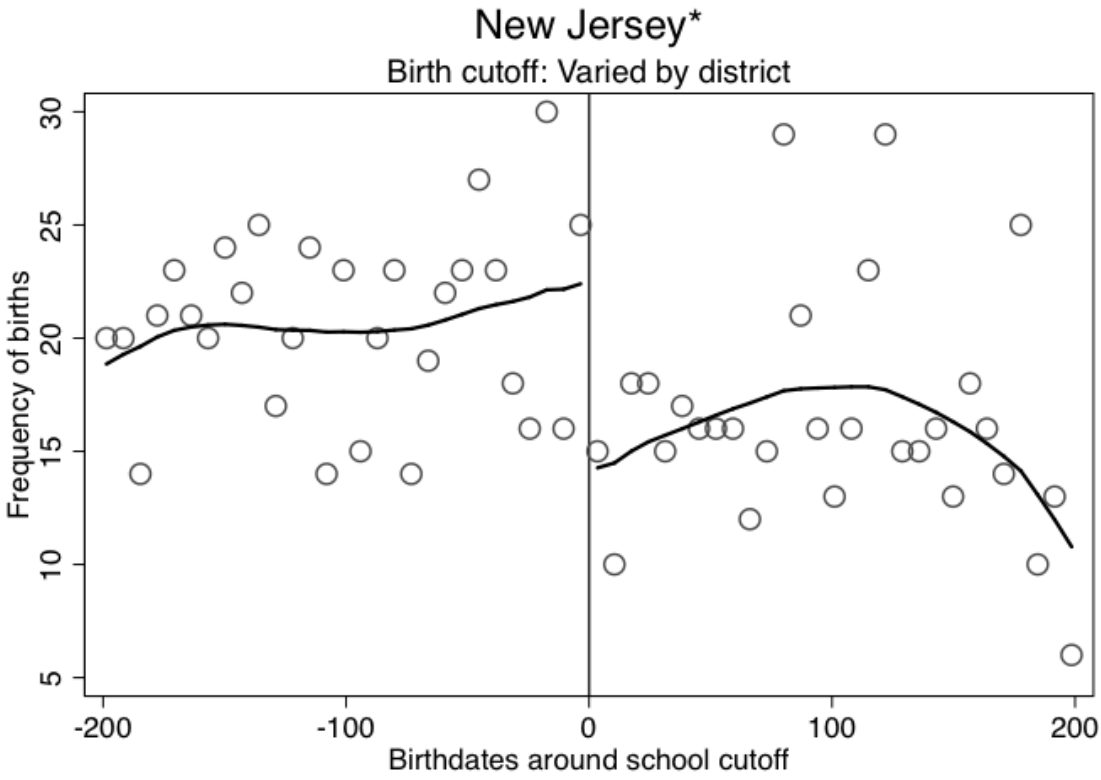

Figure 6: Smoothed frequency of birthdates relative to the cut-off in the New Jersey pre-K sample.

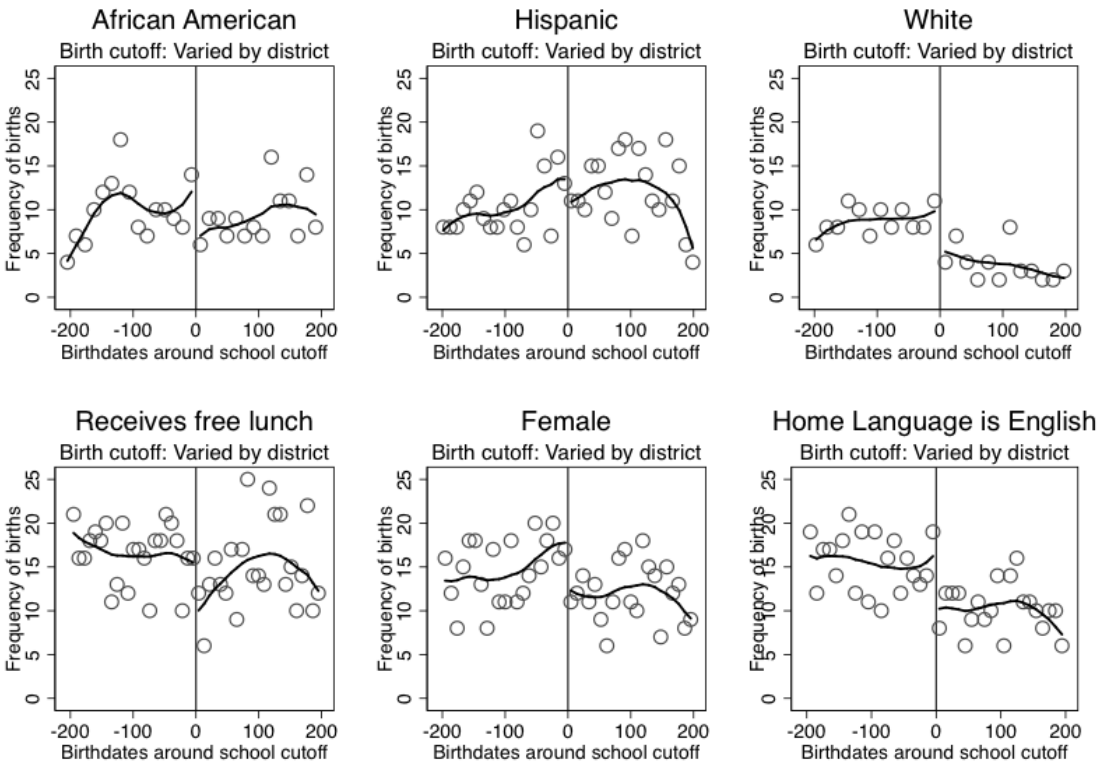

Figure 7: Smoothed frequency of birth dates by subgroups in the New Jersey pre-K sample.

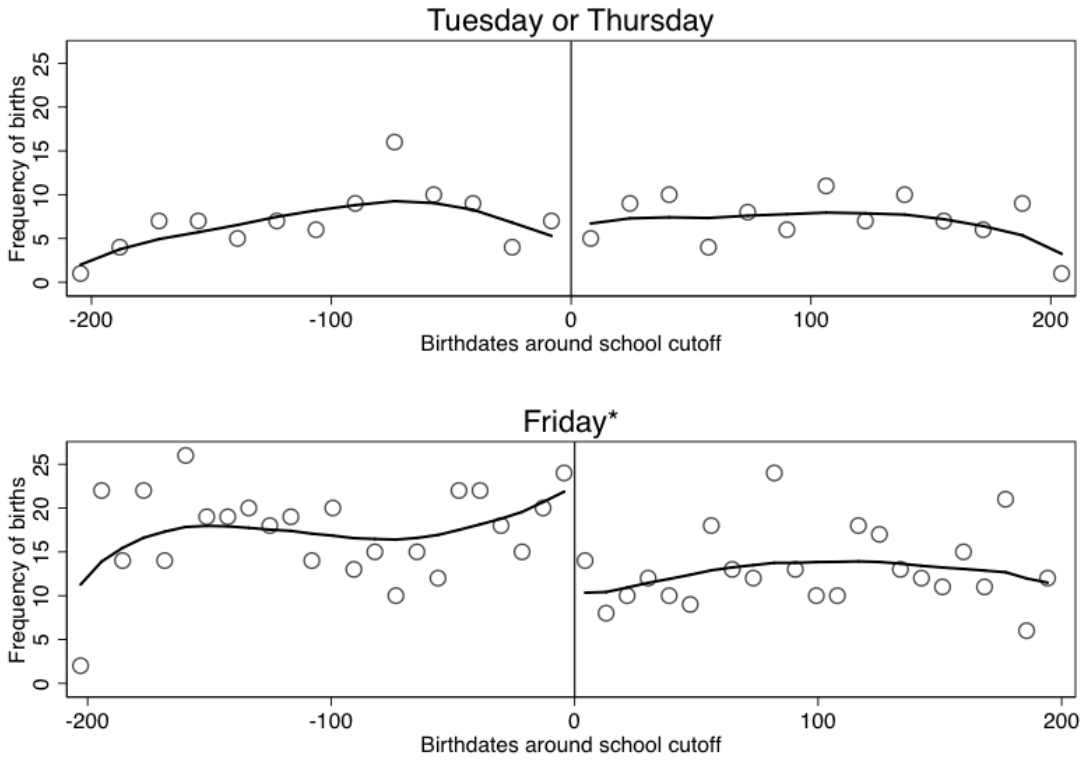

Figure 8: Smoothed frequency of birth dates using Tuesday/Thursday vs Friday cutoffs in the New Jersey pre-K sample

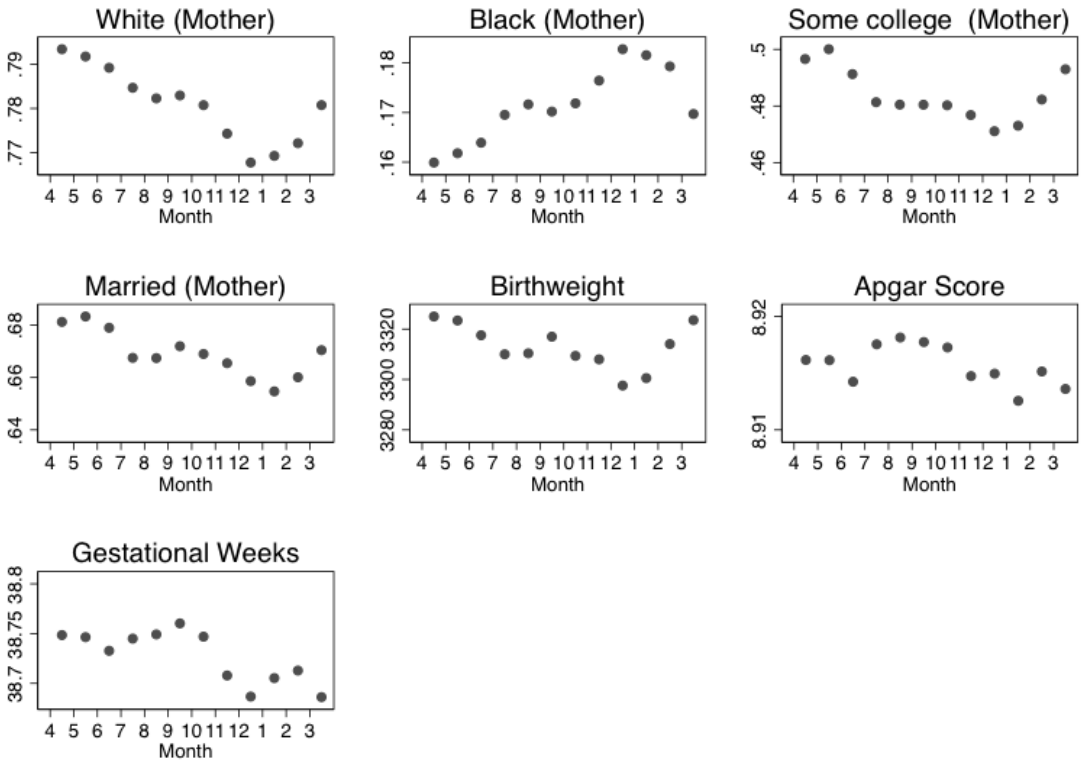

Figure 9: Population distribution of births by month (New Jersey, April 1999 to May 2000).

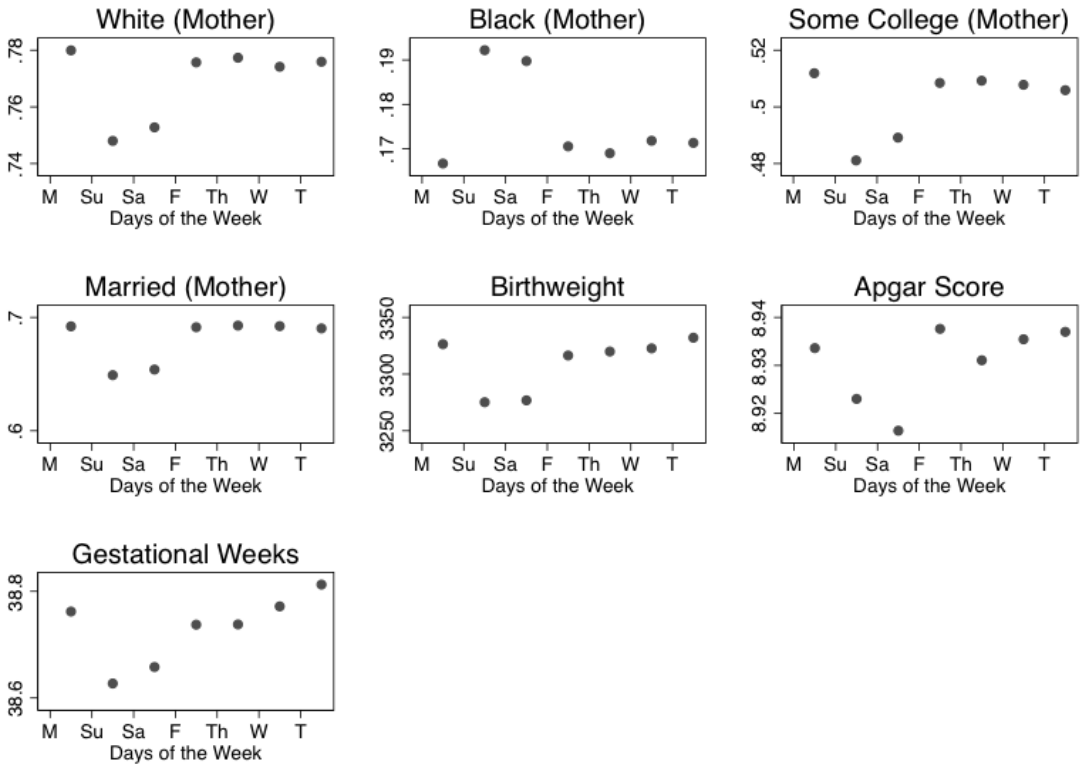

Figure 10: Population distribution of births by days of the week (New Jersey, October 1999).

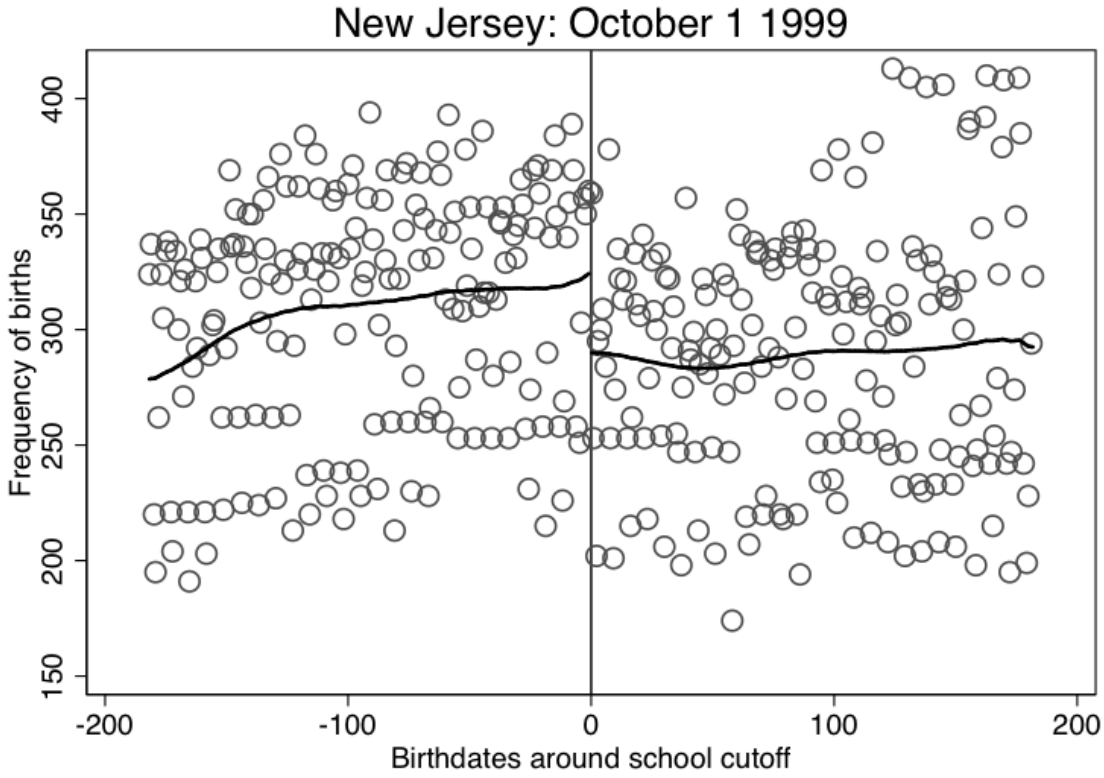

Figure 11: Manipulation test using the population distribution of New Jersey births as a reference distribution (April 1999 to May 2000).

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