NBER WORKING PAPER SERIES

HEAPING-INDUCED BIAS IN REGRESSION-DISCONTINUITY DESIGNS

Alan I. Barreca
Jason M. Lindo
Glen R. Waddell

Working Paper 17408
http://www.nber.org/papers/w17408

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2011

The authors thank Josh Angrist, Bob Breunig, David Card, Janet Currie, Todd Elder, Bill Evans, David Figlio, Melanie Guldi, Hilary Hoynes, Wilbert van der Klaauw, Thomas Lemieux, Justin McCrary, Doug Miller, Marianne Page, Heather Royer, Larry Singell, Ann Huff Stevens, Jim Ziliak, seminar participants at the University of Kentucky, and conference participants at the 2011 Public Policy and Economics of the Family Conference at Mount Holyoke College, the 2011 SOLE Meetings, the 2011 NBER's Children's Program Meetings, and the 2011 Labour Econometrics Workshop at the University of Sydney for their comments and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Heaping-Induced Bias in Regression-Discontinuity Designs
Alan I. Barreca, Jason M. Lindo, and Glen R. Waddell
NBER Working Paper No. 17408
September 2011
JEL No. C14,C21,I12

ABSTRACT

This study uses Monte Carlo simulations to demonstrate that regression-discontinuity designs arrive at biased estimates when attributes related to outcomes predict heaping in the running variable. After showing that our usual diagnostics are poorly suited to identifying this type of problem, we provide alternatives. We also demonstrate how the magnitude and direction of the bias varies with bandwidth choice and the location of the data heaps relative to the treatment threshold. Finally, we discuss approaches to correcting for this type of problem before considering these issues in several non-simulated environments.

Alan I. Barreca
206 Tilton Hall
Tulane University
New Orleans, LA
70118
abarreca@tulane.edu

Jason M. Lindo
Department of Economics
University of Oregon
1285 University of Oregon
Eugene, OR 97403
and NBER
jlindo@uoregon.edu

Glen R. Waddell
Department of Economics
University of Oregon
1285 University of Oregon
Eugene, OR 97403
waddell@uoregon.edu

1 Introduction

Empirical researchers have witnessed a resurgence in the use of regression-discontinuity (RD) designs since the late 1990s. This approach to evaluating causal effects is often characterized as superior to all other non-experimental identification strategies (Cook 2008; Lee and Lemieux 2010) as RD designs usually entail perfect knowledge of the selection process and require comparatively weak assumptions (Hahn, Todd, and van der Klaauw 2001; Lee 2008). This view is supported by several studies that have shown that RD designs and experimental studies produce similar estimates.[1] RD designs also offer appealing intuition—so long as characteristics related to outcomes are smooth through the treatment threshold, we can reasonably attribute differences in outcomes across the threshold to the treatment. In this paper, we discuss the appropriateness of this "smoothness assumption" in the presence of heaping.

For a wide variety of reasons, heaping is common in many types of data. For example, we often observe heaping when data are self-reported (e.g., income, age, height), when tools with limited precision are used for measurement (e.g., birth weight, pollution, rainfall), and when continuous data are rounded or otherwise discretized (e.g., letter grades, grade point averages). Heaping also occurs as a matter of practice, such as with work hours (e.g., 40 hours per week, eight hours per day) and retirement ages (e.g., 62 and 65). While ignoring heaping might often be innocuous, in this paper we show that doing so can have serious consequences. In particular, in RD designs, estimates are likely to be biased if attributes related to the outcomes of interest predict heaping in the running variable.

We illustrate this issue with a series of simulation exercises that consider estimating the most common of sharp-RD models,

$$Y_i = \alpha_0 + \alpha_1 1(R_i \geq c) + \alpha_2 R_i + \alpha_3 R_i 1(R_i \geq c) + \epsilon_i, \tag{1}$$

[1] See Aiken *et al.* (1998), Buddelmeyer and Skoufias (2003), Black, Galdo, and Smith (2005), Cook and Wong (2008), Berk *et al.* (2010), and Shadish *et al.* (2011) who describe within-study comparisons similar to LaLonde (1986).

where R_i is the running variable, observations with $R_i \geq c$ are treated, and ϵ_i is a random error term. As usual, this model is motivated as measuring the local average treatment effect by estimating the difference in the conditional expectations of Y_i on each side of the treatment threshold,

$$\lim_{r \uparrow c} \mathbb{E}[Y_i \mid R_i = r] - \lim_{r \downarrow c} \mathbb{E}[Y_i \mid R_i = r]. \tag{2}$$

The thought experiment embedded in this identification strategy is that in the limit as we approach the threshold, c, the treatment group and comparison group are identical in their underlying characteristics. Thus, the differences in outcomes between the two groups can be attributed to treatment.

Our primary simulation exercise supposes that the treatment cutoff occurs at zero ($c = 0$), that there is no treatment effect ($\alpha_1 = 0$), and that the running variable is unrelated to the outcome ($\alpha_2 = \alpha_3 = 0$). We introduce non-random heaping by having the expected value of Y_i vary across two data types, continuous and heaped, where continuous types have R_i randomly drawn from $[-100, 100]$ and heaped types have R_i randomly drawn from $\{-100, -90, \ldots, 100\}$. As such, we have a simple data generating process in which an attribute that predicts heaping in the running variable (type) is also related to outcomes. In this stripped-down example, we show that estimating Equation (1) will arrive at biased estimates and that the usual diagnostics are not well suited to identifying this type of problem. We also show that this bias increases when one follows the recommended practice of shrinking the bandwidth as the sample size grows (Porter 2003; Imbens and Lemieux 2008), as smaller bandwidths lead to a comparison of primarily heaped types on one side of the threshold and continuous types on the other. Furthermore, we show that non-random heaping introduces bias even if a data heap does not fall near the treatment threshold. On a brighter note, we offer alternative approaches to considering the underlying nature of the data and to accommodating non-random heaping.

To explore how non-random heaping can impair estimation in settings beyond our simple data generating process (DGP), we examine several alternative DGPs which allow the con-

tinuous and heaped types to have different means, different slopes, and different treatment effects. We consider several approaches to addressing the bias in each DGP and come to the following conclusions:

1. "Donut-RD" estimates, i.e., dropping observations at data heaps, always provides unbiased estimates of the treatment effect for continuous types. The opposite can be done to obtain unbiased estimates for heaped types, although this approach limits the extent to which one can shrink the bandwidth.

2. Allowing separate intercepts for heaped and continuous data, i.e., controlling for the underlying data type with an indicator variable, brings estimated effects closer to the average effect across the two data types (than standard-RD estimates) but does not yield estimates that are insensitive to the chosen bandwidth.

3. Allowing separate trends for heaped and continuous data, in addition to separate intercepts, brings estimated effects even closer to the average effect across the two data types but still does not yield estimates that are insensitive to the chosen bandwidth.

After conducting our simulation exercise we explore the potential for non-random heaping to bias RD-based estimates in several non-simulated environments. We begin by considering the use of birth weight as a running variable, as in Almond, Doyle, Kowalski, and Williams (2010) who estimate the effect of very-low-birth-weight classification, i.e., birth weight strictly less than 1500 grams, on infant mortality. Whereas Barreca, Guldi, Lindo, and Waddell (forthcoming) show that the estimated effect is sensitive to the treatment of observations bunched around the the 1500-gram threshold, here we consider the mechanisms leading to this bias and argue that there is a systemic problem associated with the use of birth weight as a running variable. In particular, non-random heaping is likely to lead to bias no matter what threshold one considers. Second, we show that mother's reported day of birth, used as a running variable in McCrary and Royer (2011) to estimate the effect of maternal education on fertility and infant health, also exhibits non-random heaping. Third,

we show that there is non-random heaping in income and hours worked in the Panel Study of Income Dynamics (PSID). While clearly not an exhaustive consideration of the existing RD literature or of data where heaping is evident, we provide these examples to motivate empirical researchers to consider heaping in most any exercise.

The rest of the paper is organized as follows. In Section 2, we perform a simulation exercise to demonstrate the general issue and explore potential solutions. In Section 3, we consider the problem in several non-simulated environments. We return to the general problem and the lessons learned in Section 4.

2 Simulation Exercise

2.1 Baseline Data-Generating Process and Results

As described in the introduction, we begin by considering the case in which the treatment cutoff occurs at zero ($c = 0$), there is no treatment effect ($\alpha_1 = 0$), the running variable is not related to the outcome ($\alpha_2 = \alpha_3 = 0$), and ϵ_i is drawn from $N(0, 1)$. Further, we define the sample such that 80 percent of the data have R_i drawn from a continuous uniform distribution on the interval $[-100, 100]$ (continuous types) and 20 percent of the data have R_i drawn from a discrete uniform distribution on integers $-100, -90, \ldots, 100$ (heaped types). Although one could think of the heaped types as having some underlying "true" R_i^* that is observed with error, this is not necessary for a natural interpretation since heaping can arise for a variety of reasons unrelated to measurement error. Nevertheless, it is important to note that we focus on the case in which R_i is the variable used for treatment assignment and, as such, observations are correctly classified as treated when $R_i > c$. In any case, besides having R_i drawn from different distributions, the only difference between continuous types and heaped types is their mean: continuous types have a mean of zero and heaped types have a mean of 0.5. In future sections, we refer to this data-generating process as DGP-1. All simulation exercises are based on 1,000 replications of randomly drawn samples of 10,000

observations.

These data are summarized in panels A and B of Figure 1. Panel A, which plots the distribution of the data using one-unit bins, shows the data heaps at multiples of ten. Panel B, which plots mean outcomes within the same one-unit bins with separate symbols for bins that correspond to multiples of ten, makes it clear that the means are systematically higher at data heaps.[2]

To make this example concrete, one can think of estimating the effect of free school lunches—typically offered to children in households with income below some set percentage of the poverty line—on the number of absences per week. The running variable could then be thought of as the difference between the poverty line and family income, with treatment provided when the poverty line (weakly) exceeds reported income ($R_i \geq 0$). In this example there may be heterogeneity in how individuals report their incomes—some individuals (continuous types) may report in exact amounts whereas others (heaped types) may report their incomes in tens of thousands of dollars.[3] Further, supposing that continuous types are expected to be absent zero days per week regardless of whether they are given free lunch and heaped types are expected to be absent 0.5 days per week regardless of whether they are given free lunch, then we would expect to see a mean plot similar to that of Panel B of Figure 1. That is, we have a setting in which treatment (free school lunch) has no impact on the outcome (absences). However, as we show below, the non-random nature of the heaping will cause the standard-RD estimated effects to go awry.

Continuing with DGP-1, Panel C of Figure 1 estimates the treatment effect using the standard-RD approach (Equation 1) and bandwidths ranging from one through one hundred.[4] There are two main features of this figure. First, despite the fact that there is no

[2]These plots are based on a single replication with 10,000 observations.

[3]Motivating this thought experiment, in Section 3.3 we demonstrate that there is systematic heterogeneity in how individuals report income levels, with white individuals being less likely to report incomes in thousands of dollars.

[4]Rejection rates assuming the errors are independent and identically distributed (iid) are shown in the first column of Figure A1, Panel C. Similarly, rejection rates with standard error estimates clustered on the running variable are shown in Figure A2.

treatment effect, the estimated treatment effect is always positive. Second, the bias increases as we decrease the bandwidth.[5]

When one considers the motivation for the empirical strategy, which aims at estimating the difference in the conditional expectation as we approach the treatment cutoff from each side (Equation 2), it is clear what the problem is. The identifying assumption is violated because the composition of the sample is not smooth around the cutoff. As we approach the treatment threshold from the control side, the data consist solely of continuous types in the limit. In contrast, as we approach the treatment threshold from the treatment side, the data consist solely of heaped types in the limit. Thus, even though there is no treatment effect, the expected outcome changes as we cross the treatment threshold because of the abrupt composition change induced by the data heap.

This raises several important questions which we will address in turn. How can we identify this type of problem? Is the problem specific to circumstances in which a non-random data heap falls immediately to one side of a treatment threshold? What if the heaped data has a different slope or a non-zero treatment effect? Finally, how can we address the problem once it has been diagnosed?

2.2 Diagnosing the Problem

2.2.1 Standard RD-Validation Checks are Insufficient

It is well established that practitioners should check that observable characteristics and the distribution of the running variable are smooth through the threshold. When either is not smooth through the threshold, it is usually taken as evidence that there is manipulation of the running variable (McCrary 2008). Specifically, if there is excess mass on one side of the threshold, it suggests that individuals may be engaging in strategic behavior in order to gain favorable treatment. In addition, if certain "types" are disproportionately observed on one

[5]This evidence highlights the usefulness of comparing estimates at various bandwidth levels, as proposed in van der Klaauw (2008).

side of the threshold, it suggests that there may be systematic differences in how successfully different types of individuals can manipulate the running variable in order to gain access to treatment. Any evidence of this kind is cause for concern because it suggests that the composition changes across the treatment threshold and thereby threatens identification.

Given that the problem shown in the previous section is composition bias arising from points in the distribution with excess mass, one might anticipate that our existing tools would be well suited to raising a red flag when this issue exists. As it turns out, this is not the case.

Panel A of Figure 2 shows the results of tests for a discontinuity in the distribution. In particular, data have been pooled into one-unit bins so that the frequency in each bin can be used as an outcome variable in order to estimate whether there is a discontinuity at the threshold. The set of discontinuity estimates, based on bandwidths ranging from one to one hundred, shows what one would probably expect. The estimated discontinuity is greatest when the bandwidth is small since a small bandwidth gives greater proportional weight to the data heap that falls immediately to the right of the cutoff. However, this figure also shows that we never reject zero at the five-percent level, with the exception of estimates derived from bandwidths between five and ten. As such, if the rule of thumb is that we only worry when the estimated discontinuity in the distribution is statistically significant, this diagnostic does not reliably produce the red flag one would hope for.

Given that it is obvious to the eye that there is a large data heap on the treatment side of the threshold (Figure 1, Panel A), it may be surprising that this test performs so poorly. However, this test, as described in McCrary (2008), is not meant to identify data heaps. It is meant to identify circumstances in which there is manipulation of the running variable. When there is manipulation of the running variable, we usually expect to see a dip in the distribution on one side of the cutoff as individuals close to the threshold exert effort to get to their "preferred" side. This type of behavior will produce a distribution that is qualitatively different from simply having a data heap on one side of the threshold. In particular, it will

cause there to be reduced mass on the undesirable side of the threshold at many values of the running variable, and, to the extent to which individuals overshoot the treatment threshold, excess mass on the desirable side of the threshold at many values of the running variable. In other words, this behavior will produce more of a *shift* in the distribution, which the McCrary (2008) test is well suited to identifying. In contrast, as shown above, this test is not well suited to identifying a *blip* in the distribution caused by heaping.[6]

Panel B of Figure 2 investigates the extent to which we might be able to diagnose the problem by testing whether observable characteristics are smooth through the threshold. The first graph of Panel B takes the usual RD estimation approach (Equation 1) but uses an indicator for being a heaped type as the left-hand-side variable.[7] Not surprisingly, the estimated discontinuity is greatest at small bandwidths. Rejection rates are reported in the middle graph in Panel B, which shows that we almost always reject zero if we assume the errors are independent and identically distributed (iid). To address the group structure induced by specification errors, last graph of Panel B reports rejection rates using standard error estimates clustered on the running variable, as recommended in Lee and Card (2007). Unfortunately, this practice, which has become standard, substantially reduces the likelihood that we reject zero. While we can reject zero as the bandwidth approaches zero, across most bandwidths we *never* reject zero. Of course, pure noise on the left-hand side implies a rejection rate of five percent.

These results paint a pessimistic picture for multiple reasons. First, in practice, where one is likely to follow the recommended procedure of clustering standard-error estimates, this diagnostic test only identifies the problem as the bandwidth approaches zero. Second, the test performs even worse (i.e., rejects zero less often) at smaller bandwidths if the researcher

[6]Another way of describing this issue is that the test is meant to identify low-frequency changes, while heaping results in very-high-frequency changes.

[7]Although unlikely in practice, it is instructive to suppose that the researcher can identify continuous types and heaped types to illustrate how well the diagnostic test performs under ideal circumstances. This would be analogous, for example, to knowing which observations of income are approximations, because some individuals reporting incomes in tens of thousands will be providing an approximation (heaped types) and others reporting the same incomes will actually be giving exact amounts (continuous types).

does not observe type but instead observes only a proxy for type.[8] Finally, this approach will perform worse still if there is not a data heap exactly at the threshold. This is unfortunate, as will become clear in Section 2.3, since data heaps further from the threshold also bias estimated treatment effects.

2.2.2 Supplementary Approaches to Identifying Non-random Heaping

Although the conventional specification checks are well suited to diagnosing strategic manipulation of the running variable, the results above demonstrate that additional diagnostics are required to identify non-random heaping. In this section, we introduce two such diagnostics.

First, while RD papers are inclined to produce mean plots as standard practice, the level of aggregation is typically such that non-random heaping can be hidden. Moreover, not visually distinguishing heaped data points from others can lead one to mistake the heap points for noise rather than the systematic outliers that they are. As a simple remedy, however, one can show disaggregated mean plots that clearly distinguish heaped data from non-heaped data, as in Panel B of Figure 1.[9] Of course, a necessary pre-requisite to this approach is knowledge of where the heaps fall, which can be revealed through a disaggregated histogram, as in Panel A of Figure 1.

Although this approach is useful for visual inspection of the data, a second, more-rigorous approach is warranted when systematic differences between heaped data and non-heaped data are less obvious. As a more formal way of testing whether a given data heap Z systematically deviates from its surrounding non-heaped data, one can estimate

$$X_i = \gamma_0 + \gamma_1 1(Z_i = Z) + \gamma_2(R_i - Z) + u_i, \tag{3}$$

using the data at heap Z itself in addition to the continuous data within some bandwidth

[8] For example, in Section 3.1 we show that measures of low-socioeconomic status are imperfect predictors of heaping.

[9] We recommend this type of plot as a complement to more-aggregated mean plots rather than a substitute. For example, more-aggregated mean plots are more useful when trying to discern what functional form should be used in estimation and whether or not there is a treatment effect.

around Z. Essentially, this regression equation estimates the extent to which characteristics "jump" off of the regression line predicted by the continuous data. If γ_1 is significant then one can conclude that the composition changes abruptly at heap Z.[10]

2.3 What if Heaping Occurs Away from the Threshold?

In the previous section, we described a scenario where a data heap coincided closely with the treatment threshold. In this section, and before we move on to alternative data-generating processes altogether, we examine the extent to which non-random heaping leads to bias when a heap does not fall immediately to one side of the threshold. In particular, Panel A of Figure 3 shows the estimated treatment effect as we change the location of the heap relative to the cutoff, while rejection rates are shown in Panel B. In order to focus on the influence of a single heap, we adopt a bandwidth of five throughout this exercise.

As before, when the data heap is adjacent to the cutoff on the right, the estimated treatment effect is biased upwards. Conversely, when the data heap is adjacent to the cutoff on the left, the estimated treatment effect is biased downwards. While this is rather obvious, what happens as we move the heap away from the cutoff is perhaps surprising. Specifically, the bias does not converge to zero as we move the heap away from the cutoff. Instead, the bias goes to zero *and then changes sign*. This results from the influence of the heaped types on the slope terms. For example, consider the case in which the data heap (with mean 0.5 whereas all other data are mean zero) is placed five units to the right of the treatment threshold and the bandwidth is five. In such a case, the best-fitting regression line through the data on the right (treatment) side of the cutoff will have a positive slope and a negative intercept. In contrast, the best-fitting regression line through the data on the left (control) side of the cutoff will have a slope and intercept of zero. Thus, we arrive at an estimated effect that is negative. The reverse holds, of course, when we consider the case in which the

[10]This test is somewhat lacking in information when performed on our simulated data where, by construction, we can perfectly identify changes in the attribute (type) at heaps. In Section 3.1, we show these types of estimates applied to "real" data where attributes are not perfectly related to heaping (Figure 9).

data heap is five units to the left of the treatment threshold, which yields an estimated effect that is positive. Recall that, in all cases, the treatment effect is zero.

2.4 Alternative Data-Generating Processes

In order to demonstrate other issues related to non-random heaping, in this section we explore several alternative DGPs. We begin with two additional DGPs in which there is no treatment effect, and then consider DGPs in which there is a treatment effect. We summarize each of six DGPs in Figure 4.

The next DGP (DGP-2) is the same as the baseline DGP-1 except that the continuous types and heaped types have different slopes instead of different means. In particular, in DGP-2 the mean is zero for both groups, the slope is zero for continuous types, and the slope is 0.01 for heaped types. These parameters are summarized in the second column of Figure 4, in Panel A, while the means are shown in Panel B. Panel C shows the estimated treatment effects for varying bandwidths. Again, although there is no treatment effect for continuous or heaped types, the estimates tend to suggest that the local average treatment effect is not zero. The estimates display a sawtooth pattern, dropping to less than zero each time an additional pair of data heaps is added to the analysis sample and then climbing above zero again as additional continuous data is added. As in DGP-1, the bias becomes less severe at larger bandwidths, although this may not be obvious to the eye.

To understand the sawtooth pattern first exhibited in DGP-2, before continuing to other data-generating processes, Figure 5 plots the regression lines using selected bandwidths. The short-dashed lines are based on a bandwidth of 10, where the data include three heap points, $R = \{-10, 0, 10\}$. These lines show that the non-random heaping captured in DGP-2 leads to an estimate that is negatively biased. In particular, the heap at $R = -10$ has two effects on the regression line on the left side of the threshold. First, it causes the regression line to shift down because it pulls down the center of mass.[11] Second, it induces a positive slope in

[11] Recall that a regression line always runs through $(\bar{x}, \bar{y})$.

order to bring the the regression line closer to the heaped data at the edge of the bandwidth. As it turns out, the slope is large enough that the regression line crosses zero from below, which results in a positive expected value approaching the treatment threshold from the left. The heap at $R = 10$ has similar effects on the regression line on the right side of the threshold—it shifts the regression line up and induces a positive slope such that the expected value is negative approaching the treatment threshold from the right. As such, approaching the threshold from each side, we arrive at a negative difference in expected value.

The dash-and-dot line in Figure 5 uses a bandwidth of 18 to demonstrate how the same DGP can arrive at positive estimates. Again, on both the left and right sides of the treatment threshold, the sum of squared errors are minimized by a positively-sloped regression line. However, with more continuous data, including a sizable share to the left of the data heap at $R = -10$ and to the right of the data heap at $R = 10$, the magnitude of the slope is much smaller. As a result, neither regression line, on the left side or the right side of the threshold, crosses zero. Thus, we have a negative expected value approaching the treatment threshold from the left and a positive expected value approaching the treatment threshold from the right, i.e., a positive estimate of the treatment effect.

Last, the solid line in Figure 5 plots the regression lines using a bandwidth of 20. Here, it is important to keep in mind that the increase in the bandwidth has introduced data heaps at $R = -20$ and $R = 20$ to the analysis. Not surprisingly, it shares a lot in common with the short-dashed line that plotted regression lines using a bandwidth of 10. In particular, the heaps at the boundary of the bandwidth influence the slope parameters such that the regression lines cross zero. As such, we again find a negative estimate of the treatment effect when the true effect is zero.

As shown in the second column of Panel B in Figure 4, this phenomena occurs in a systematic fashion as we change the bandwidth. Each time a new set of heaps is introduced, the slope estimate becomes sharply positive, which leads the regression lines on each side of the cutoff to pass through zero, which leads to negative estimates of the treatment effect. As we

increase the bandwidth beyond a set of heaps, however, the slope terms shrink in magnitude, the regression lines no longer pass through zero, and we arrive at positive estimates of the treatment effect. The process repeats again when the increase in bandwidth introduces a new set of heaps.

This sawtooth pattern will remain evident as we consider additional DGPs. For example, the third column of Figure 4 considers DGP-3, which combines elements of both DGP-1 and DGP-2. In particular, the parameters for the continuous data remain the same (set at zero) whereas the heaped data have a higher mean (0.5) and a higher slope (0.01). Not surprisingly, the estimated local average treatment effects exhibit a positive bias that grows exponentially as the bandwidth is reduced (as in DGP-1) and a sawtooth pattern (as in DGP-2).

We now turn to DGPs in which there is a treatment effect for heaped types while maintaining the same parameters for the continuous types. Specifically, for heaped types in DGP-4, we set the control mean equal to zero and set the treatment effect to 0.5. Given that there is no treatment effect for continuous types, who comprise 80 percent of the data, the average treatment effect is 0.1. The set of estimates in Panel C shows that the estimated treatment effects converge on the average treatment effect as the bandwidth grows. As the bandwidth shrinks to zero, however, the estimated treatment effects approach 0.5. This occurs because the expected outcome immediately to the right of the treatment threshold is 0.5 (based on the heaped that is data nearest the threshold) whereas the expected outcome immediately to the left of the treatment threshold is zero (based on the continuous data that is nearest the threshold).

It is important to note that estimates will not generally converge on the true treatment effect for heaped data as we shrink the bandwidth. We demonstrate this fact in DGP-5, which simply raises the mean value for the heaped data by 0.5. With this DGP, the estimates again converge to 0.1 as the bandwidth increases. However, as we shrink the bandwidth, the estimates converge on one, an estimate that should not be interpreted as any sort of true treatment effect. We obtain this estimate because the expected outcome immediately to the

right of the treatment threshold is one (based on the heaped that is nearest the threshold) whereas the expected outcome immediately to the left of the treatment threshold is zero (based on the continuous data that is nearest the threshold).

Finally, with DGP-6, we consider a case in which the heaped data has a higher control mean than the continuous data (0.5 versus zero), has a treatment effect where the continuous data does not (0.5 versus zero), and has a positive slope where the continuous data does not (0.1 versus zero). As in the other DGPs in which the heaped data had a different slope from the continuous data (DGP-2 and DGP-3), the estimates based on DGP-6 exhibit a prominent sawtooth pattern. In addition, as in DGP-5, which has the same means, the estimates converge to 0.1 as we grow the bandwidth.

Panel C of Appendix Table A1 shows rejection rates at the five-percent level based on the assumption that errors are independent and identically distributed. Similarly, Panel C of Appendix Table A2 shows rejection rates when we cluster the standard error estimates on the running variable. These figures show that we always over-reject zero for all of the DGPs when we assume iid errors. Clustering the standard errors on the running variable instead leads to over-rejection at small bandwidths and under-rejection at large bandwidths.

2.5 Addressing Non-Random Heaping

In this section, we explore three approaches to addressing the bias induced by non-random heaping. To begin, we consider a donut-RD, of sorts, in which we simply drop the heaped data from the analysis. Estimates based on this approach are shown in Panel D of Figure 4. For all DGPs and all bandwidths, this approach leads to an unbiased estimate of the treatment effect for continuous types, which is zero.

In practice, there are several tradeoffs to weigh when considering a donut-RD approach. First, data heaps may comprise a large share of the data. Second, one might be interested in knowing about the treatment effect for types who tend to be observed at data heaps. In either case, it might make sense to instead focus only on the heaped data in estimation.

The cost of this approach, however, is that it limits how close one can get to the treatment threshold to obtain estimates. For example, when there are data heaps at multiples of ten units, 20 is the smallest bandwidth one could use to estimate Equation 1 since it requires at least two observations on each side of the threshold. Of course, one would likely want to use a much larger bandwidth to avoid problems associated with having too few clusters.

Given the advantages offered by focusing on the continuous data—unbiased estimates and the ability to shrink the bandwidth towards zero—the donut approach would seem to be an integral part of any RD-based investigation in the presence of heaping. Although less convincing since it requires larger bandwidths, it may also be reasonable to separately estimate the effects focusing solely on the heaped data.

At the same time, one may be interested in continuing to estimate the treatment effect using a pooled model. For example, one may take this approach because their sample size is limited, making separate analyses of the continuous data difficult, or because they are interested in estimating an average treatment effect for all types using a single regression.

Panel E of Table 4 takes this approach by adding an indicator variable for being at a data heap to the standard RD model. These estimates clearly show that adding this control variable is no panacea. While it does fully remove the bias from estimates in DGP-1, in which the only difference between the continuous and heaped data is their mean, it does not fully remove the bias from estimates based on other DGPs.

Panel F takes a more flexible approach to estimation based on the pooled data, allowing separate intercepts and trends for the heaped data. This approach does better than simply allowing a different intercept, as in Panel E. It removes the bias in every case in which the continuous and heaped data have different means or slopes while having the same treatment effect (DGP-1, DGP-2, and DGP-3). In addition, for the DGPs in which the heaped data have a different treatment effect than the continuous data (DGP-4, DGP-5, and DGP-6), this approach brings the estimates closer to the average treatment effect, although it does not retrieve unbiased estimates of the average treatment effect.

It may come as a surprise that this approach does not lead to a simple weighted average of the treatment effects one would obtain if one estimated the discontinuity separately for each data type.[12] Further investigation of alternative DGPs (not reported) reveals that this issue is caused by the presence of heterogeneous treatment effects rather than non-random heaping. We see this as an important area for future research that is beyond the scope of this paper.[13]

3 Non-Simulated Examples

In this section, we first focus on birth weight, which has previously been used as a running variable to identify the effects of hospital care on infant health. We show that birth weight data exhibits non-random heaping that biases RD estimates, that the usual diagnostics fail to detect the problem whereas our proposed diagnostics prove useful, and that the approaches that were effective at reducing the bias in the simulation are also effective in this context. We then turn our attention to date of birth, which has previously been used as a running variable to identify the effects of maternal education on fertility and infant health. We show that these data also exhibit non-random heaping which could lead to bias. Last, using the PSID, we show that non-random heaping is present in two frequently used variables: work hours and earnings.

3.1 Birth Weight as a Running Variable

3.1.1 Background

Because some hospitals use birth-weight cutoffs as part of their criteria for determining medical care, a natural way of measuring the returns to such care is to use a RD design

[12] It came as a surprise to us, writing just the opposite in earlier drafts before conducting the formal simulation exercises described here.

[13] We explore this issue further in work in progress. For example, if one considers two continuous data types that have different treatment effects one will not obtain estimates of the average treatment effect that are insensitive to the chosen bandwidth.

with birth weight as the running variable. Almond, Doyle, Kowalski, and Williams (2010), hereafter ADKW, take this approach in order to consider the effect of very-low-birth-weight-classification, i.e., having a measured birth weight strictly less than 1500 grams, on infant mortality. Whereas Barreca, Guldi, Lindo, and Waddell (forthcoming) shows that the estimated effect is sensitive to the treatment of observations bunched around the 1500-gram threshold, here we consider the use of birth weight as a running variable more broadly, which sheds light on *why* estimates are sensitive to the treatment of observations bunched around the threshold.[14] In doing so, we demonstrate that the problems are inherent throughout the birth weight distribution and not particular to a specific setting or margin of interest.

To begin, consider that birth weights can be measured using a hanging scale, a balance scale, or a digital scale, each of them rated in terms of their resolution. Modern digital scales marketed as "neonatal scales" tend to have resolutions of 1 gram, 2 grams, or 5 grams. Products marketed as "digital baby scales" tend to have resolutions of 5 grams, 10 grams, or 20 grams. Mechanical baby scales tend to have resolutions between 10 grams and 200 grams. Birth weights are also frequently measured in ounces, with ounce scales varying in resolution from 0.1 ounces to four ounces. Because not all hospitals have high performance neonatal scales, especially going back in time, a certain amount of heaping at round numbers is to be expected.

In the discussion of the simulation exercise above we recommended that researchers produce disaggregated histograms for the running variable. We do so for birth weights in Figure 6.[15] As also noted in ADKW, this figure clearly reveals heaping at 100-gram and

[14] In contrast to the estimated-reduced-form effect of very-low-birth-weight classification on mortality, Almond, Doyle, Kowalski, and Williams (forthcoming) show that it turns out that the instrumental-variables estimate of the effect of hospital spending (triggered by very-low-birth-weight classification) is not very sensitive to the treatment of observations around the threshold when one focuses on selected states.

[15] We use identical data to ADKW throughout this section, Vital Statistics Linked Birth and Infant Death Data from 1983–1991 and 1995–2002; linked files are not available for 1992–1994. These data combine information available on an infant's birth certificate with information on the death certificate for individuals less than one year old at the time of death. As such, the data provides information on the infant, the infant's health at birth, the infant's death (where applicable), the family background of the infant, the geographic location of birth, and maternal health and behavior during pregnancy. For information on our sample construction, see Almond, Doyle, Kowalski, and Williams (2010).

ounce multiples, with the latter being most dramatic. Although we focus on these heaps throughout the remainder of this section to elucidate conceptual issues involving non-random heaping, a more complete analysis of the use of birth weights as a running variable would need to consider heaps at even smaller intervals (e.g., 50-grams, half-ounces). In any case, to the extent to which any of the observed heaping can be predicted by attributes related to mortality, our simulations imply that standard-RD estimates are likely to be biased.

In considering the potential for heaping to be systematic in a way that is relevant to the research question, we first note that scale prices are strongly related to scale resolutions. Today, the least-expensive scales cost under one hundred dollars while the most expensive cost approximately two thousand dollars. For this reason, it is reasonable to expect more-precise birth weight measurements at hospitals with greater resources, or at hospitals that tend to serve more-affluent patients.[16] That is, one might anticipate that the heaping evident in Figure 6 is systematic in a non-trivial way.

3.1.2 Diagnostics

ADKW noted that there was significant heaping at round-gram numbers and at gram equivalents of ounce multiples. However, they did not test whether the heaping was random. They did, of course, perform the usual specification checks to test for non-random sorting across the treatment threshold. Despite there being two obvious heaps immediately to the right of the treatment threshold at 1500 grams and 1503 grams (i.e., 53 ounces), they find that the estimated discontinuity in the distribution is not statistically significant, which is not surprising given the results of our simulation exercise.

In addition, ADKW make the rhetorical argument that there are not irregular heaps around the 1500-gram threshold of interest since the heaps are similar around 1400 and

[16]With general improvement in technology one would anticipate that measurement would appear more precise in the aggregate over time. We show that this is indeed the case in Appendix Figure A3 which also foreshadows the systematic relationship between between heaping and measures of socioeconomic status. Note that a major reason that the figure does not show smooth trends is because data is not consistently available for all states.

1600 grams. With respect to the usual concerns about non-random sorting, this argument is compelling. In particular, the usual concern is that agents might engage in strategic behavior so that they are on the side of the threshold that gives them access to favorable treatment. While this is a potential issue for the 1500-gram threshold, it is not an issue around 1400 and 1600 grams. Since we also see heaping at the 1400- and 1600- gram thresholds, it makes sense to conclude that the heaping observed at the 1500-gram threshold is "normal." The problem with this line of reasoning, however, is that *all* of the data heaps may be systematic outliers in their composition.

Panels A and B of Figure 7 replicate ADKW's analysis of covariates at the 1,500-gram cutoff along with the placebo cutoffs of $1,000, 1,100, \ldots, 3,000$ grams.[17] In particular, we use the regression described in Equation (1) and an 85-gram bandwidth to consider the extent to which there are mean shifts in child characteristics across the considered thresholds. We plot percent changes, 100 multiplied by the estimated treatment effect divided by the intercept, for greater comparability across the differing cutoffs.[18] These plots show that there are rarely statistically significant discontinuities in covariates, such as the probability that a mother is white and the probability that a mother has less than a high-school education. However, the set of estimates do reveal a distinct pattern that is informative—more often than not, the estimates suggest that those just below these cutoffs are of lower socioeconomic status. Similarly, panels C and D of Figure 7 consider Apgar scores, a measure of health taken five minutes after birth.[19] These estimates suggest that children just below 100-gram thresholds have higher average Apgar scores and a lower chance of being born with an Apgar score

[17]We note that not all of these are "true placebo cutoffs" as 1,000 grams corresponds to the extremely-low-birth-weight cutoff and 2,500 grams corresponds to the low birth weight cutoff.

[18]We bootstrap standard errors and cluster on exact grams. Each of our 500 bootstrap replications draws C observations at random, where C is the number of clusters in the full data set, and then collects all observations within each drawn cluster for the analysis.

[19]While it is conceivable that Apgar scores might be affected by treatment induced by very-low-birth-weight classification at the 1,500-gram cutoff, there is no reason to expect this to be the case at so many other 100-gram multiples. Around these other thresholds, we can say with relative confidence that in the absence of composition bias Apgar scores should vary smoothly along the birth weight distribution. Note that Apgar scores are not available for all state-years in the Vital Statistics. However, they are reported for the majority of births—approximately 75 percent.

below three (out of a ten point scale), but they are rarely significant at the 95 percent level.[20] Again, if one's attention were restricted to a single threshold, this approach would not reliably produce a red flag.

Following our recommended procedure, Figure 8 plots average child characteristics against recorded birth weights, visually differentiating heaps at 100-gram and ounce intervals. This is our first strong evidence that the heaping apparent in Figure 6 is non-random—children at the 100-gram heaps are disproportionately likely to be nonwhite (Panel A), have mothers with less than a high-school education (Panel B), and have low Apgar scores (panels C and D). In contrast, these graphs suggest that children at ounce heaps are disproportionately likely to be white, have mothers with at least a high-school education, and have high Apgar scores. However, compared to those at 100-gram heaps, it is far less clear that those at ounce heaps are outliers in their underlying characteristics, highlighting the usefulness of our second approach.

As a second approach to exploring the extent to which the composition of children changes abruptly at reporting heaps, we estimate the regression equation:

$$X_i = \gamma_0 + \gamma_1 1(BW_i = H) + \gamma_2(BW_i - H) + u_i, \tag{4}$$

for $H = \{1{,}000,\ 1{,}100,\ ...,\ 3{,}000\}$ and gram equivalents of ounce multiples, where X_i is a characteristic of individual i with birth weight BW_i. As discussed in the simulation exercise, (4) is not intended to detect a mean shift across H but, rather, the extent to which characteristics heaped at H differ from what would be expected based on surrounding observations that are not at data heaps.[21]

The results from this regression analysis confirm that child characteristics change abruptly

[20] To provide a frame of reference, 58 percent of children with Apgar scores less than three survive at least one year whereas the rate of one-year survival is 0.99 for births with Apgar scores greater than three.

[21] We note that one could allow for different trends on each side of the data heaps. We do not take this approach for two reasons. As a practical matter, testing for mean deviations at heaps approaching both sides would double the set of reported estimates and could lead to confusion. More importantly, there is generally not a good reason to expect slopes to change when crossing most data heaps.

at data heaps. Focusing on the 100-gram heaps, Panel A of Figure 9 shows estimated percent changes, γ_1/γ_0, for the probability that a mother is white, the probability that she has less than a high-school education, Apgar score, and the probability of having an Apgar score less than four out of ten. For nearly every estimate, bootstrapped standard error estimates clustered at the gram level are small enough to reject that the characteristics of children at H are on the trend line. Similarly, Panel B of Figure 9 demonstrates that those at *ounce* heaps also tend to systematically deviate from the trend based on surrounding non-ounce observations, except the more-affluent types are disproportionately more likely to have birth weights recorded in ounces. Although the estimates for each ounce heap are rarely statistically significant, it is obvious that the set of estimates is jointly significant and that the individual estimates would usually be significant with a bandwidth larger than 85 grams. In the end, where the standard validation exercises do not detect this important source of potential bias, this simple procedure proves effective.

3.1.3 Non-random Heaping, Bias, and Corrections

Given these relationships between characteristics that predict both infant mortality and heaping in the running variable, our simulation exercise suggests standard-RD estimates will be biased. To illustrate this issue, we continue to pursue estimates based on the standard-RD approach characterized by Equation (1), exploring the set of cutoffs of $1,000, 1,100, \ldots, 3,000$ grams that largely can be thought of as placebo tests. We now simply consider one-year mortality and 28-day mortality on the left-hand side.

Panel A of Figure 10 presents the estimated percent impacts on one year mortality and 28-day mortality. These figures suggest that, near any of the considered cutoffs c, children with birth weights less than c routinely have better outcomes than those with birth weights at or above c. Given that these results are largely driven by a systematically different composition of children at the 100-gram heaps that coincide with the cutoffs, the estimated effects are much larger in magnitude when one uses a narrow bandwidth. For example, with

a bandwidth of 30 grams, 42 of 42 point estimates fall below zero.[22]

Our simulation exercise considered several ways to deal with this type of composition bias. First, to the extent to which one might believe that there is only selection on observables, one may attempt to eliminate the bias by controlling for covariates.[23] Panel B of Figure 10 shows the results of this approach, controlling with fixed effects for state, year, birth order, the number of prenatal care visits, gestational length in weeks, mothers' and fathers' ages in five-year bins (with less than 15 and greater than 40 as additional categories), and controlling with indicator variables for multiple births, male, black, other race, Hispanic, the mother having less than a high-school education, the mother having a high-school education, the mother having a college education or more, and whether the mother lives in her state of birth. These estimates are qualitatively similar to those reported in Figure 7. In almost all cases, they imply that having a birth weight less than the considered cutoff reduces infant mortality. We interpret this set of estimates as evidence that this approach has not effectively dealt with the composition bias. A more general problem with this type of approach is that it is difficult to test the extent to which the added covariates improve the match between the treatment and control groups since such a test will usually entail a comparison of the covariates themselves. However, considering placebo cutoffs can be useful in this regard.

As illustrated in the simulation exercise, a more effective approach to dealing with non-random heaping is to perform a donut RD, estimating the effect after dropping observations

[22]Results are similar if one uses triangular kernel weights which also place greater emphasis on observations at 100-gram heaps. ADKW mention having considered the effects at these same placebo cutoffs, motivating the analysis as follows:

> "[A]t points in the distribution where we do not anticipate treatment differences, economically and statistically significant jumps of magnitudes similar to our VLBW treatment effects could suggest that the discontinuity we observe at 1500 grams may be due to natural variation in treatment and mortality in our data."

They do not present these results but instead report:

> "In summary, we find striking discontinuities in treatment and mortality at the VLBW threshold, but less convincing differences at other points of the distribution. These results support the validity of our main findings."

We disagree with this interpretation of the results.

[23]The only covariate available in the simulation exercise, of course, was "data type."

at data heaps. While a drawback of this method is that it cannot tell us about the treatment effect for the types who tend to be observed at data heaps, it is consistent with the usual motivation for RD. Specifically, it has the researcher focus on what might be considered a relatively-narrow sample in order to be more confident that we can identify unbiased estimates.

Panel C of Figure 10 shows the donut-RD estimated effects on infant mortality, omitting those at 100-gram and ounce heaps from the analysis. While the earlier estimates (in panels A and B) were negative for most of the placebo cutoffs, these estimates resemble the white-noise process we would anticipate in the absence of treatment effects. Thus, these results indicate that the sample restrictions we employ reduce the bias produced by the non-random heaping described above. These results also suggest that the estimated impact of very low birth weight classification is zero.

Our simulation exercise also showed that allowing separate trends for each "observation type" reduced the bias introduced by non-random heaping. Since it is clearly not possible to use this approach to deal with observations at 100-gram heaps without substantially increasing the bandwidth, we drop these observations and use this approach on the remaining data. In particular, we allow the slope and intercept terms to be different for those at ounce heaps and those not at ounce heaps.[24] Panel D of Figure 10 shows the results of this approach. These estimates do not reveal systematic mortality reductions to the left of 100-gram cutoffs which we again take as support for the usefulness of this approach.

3.2 Date of Birth as a Running Variable

A common approach to estimating the effects of education on outcomes is to use variation driven by small differences in birth timing that straddle school-entry-age cutoffs. For example, "five years old on December 1st" is a common school-entry requirement. As such, by comparing the outcomes of individuals who are born just before December 1st, who begin

[24] In doing so, we estimate seven parameters rather than the usual four which usually correspond to the constant, the discontinuity estimate, and two slope terms.

school earlier and thereby tend to obtain more years of education, to those of individuals who are born just after December 1st, one can measure the causal effect of education on outcomes.

McCrary and Royer (2011), for example, do just this in order to identify the causal effect of maternal education on fertility and infant health using restricted-use birth records from California and Texas. In the first graph of Figure 11, Panel A, we use the same California birth records as McCrary and Royer and show the distribution of mothers' reported birth dates across days of the month.[25] Although less prominent than in the birth weight example, this figure shows that there are data heaps at the beginning of each month and at multiples of five. The second graph in Panel A shows one of many indications that those at data heaps are outliers—that the mothers at these data heaps are disproportionately less likely to have used tobacco during their pregnancies. This phenomenon is not specific to tobacco use, however. Similar patterns are equally evident in mother's race, father's race, mother's education, father's education, the fraction having father's information missing, or the fraction having pregnancy complications, along with a wide array of child outcomes.[26]

It turns out that this non-random heaping is unlikely to be a serious issue for the main results presented in McCrary and Royer (2011) because their preferred bandwidth of 50 leaves their estimates relatively insensitive to the high frequency composition shifts described above.[27] At the same time, it is important to keep in mind that recommended practice would have them choose a smaller bandwidth were more data available. Our simulation

[25]The California Vital Statistics Data span 1989 through 2004. These data, obtained from the California Department of Pubic Health, contain information on the universe of births that occurred in California during this time frame. Mothers date of birth is not available in the public use version of the National Vital Statistics Natality Data. We use the same sample restrictions as McCrary and Royer (2011), limiting the sample to mothers who: were born in California between 1969 and 1987, were 23 years of age or younger at the time of birth, gave birth to her first child between 1989 and 2002, and whose education level and date of birth are reported in the data.

[26]For related reasons, the empirical findings in Dickert-Conlin and Elder (forthcoming) should also be considered in future papers that use day of birth as their running variable. In particular, they show that there are relatively few children born on weekends relative to weekdays because hospitals usually do not to schedule induced labor and cesarian sections on weekends. As such, children born without medical intervention who tend to be of relatively low socioeconomic status are disproportionately observed on weekends.

[27]With that said, this phenomenon may explain why their estimates vary a great deal when their bandwidth is less than twenty but are relatively stable at higher bandwidths.

exercise demonstrates that this practice would make the problems associated with non-random heaping more severe.

3.3 Labor-Market Outcomes in the PSID

As we described in the motivation for our simulation exercises, one might consider using income as a running variable in an RD to measure the impact of free school lunch on child outcomes. Of course, given how many policies are income-based, there are several other examples where treatment effects might be identified using an RD with income as the running variable. For example, one might consider this strategy to identify the effects of various tax incentives, the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), or the Children's Health Insurance Program (CHIP) which subsidizes health insurance for families with incomes that are marginally too high to qualify for Medicaid.

The Panel Study of Income Dynamics (PSID) would be a natural data source for such studies. It is one of the most widely used data sets in labor economics research because it provides a representative sample of the U.S., tracks individuals and their descendants from 1968 through present, and collects an immensely rich set of variables.

Focusing on nominal incomes for PSID heads of household 1968–2007, Panel B of Figure 11 shows that there is significant heaping at \$1,000 multiples and that individuals at these data heaps are outliers in their underlying characteristics. In particular, those at \$1,000 heaps are substantially less likely to be white than those with similar incomes who are not found at these data heaps.[28] As demonstrated in our previous sections, this could clearly lead to biased estimates.

Lastly, Panel C of Figure 11 uses the same data to consider annual hours of work, which could also be used as a running variable in an RD. For example, many employers only provide health insurance and other benefits to employees who work some predetermined number of

[28]For visual clarity, these graphs focus on individuals with positive incomes less than \$40,000, which is approximately equal to the 75th percentile. In addition, the histogram uses \$100 bins and the mean plot uses \$100 bins for the data that are not found at \$1,000 multiples.

hours. In these data, it is evident that there is heaping at 40-hour multiples.[29] Again, it does not appear to be random. Those at the 40-hour heaps have less education, on average, than those who work a similar number of hours who are not found at these data heaps, which would need to be addressed if hours were to be used as the running variable in an RD.

4 Discussion and Conclusion

In this paper, we have demonstrated that the RD design's smoothness assumption is inappropriate when there is non-random heaping. In particular, we have shown that RD-estimated effects are afflicted by composition bias when attributes related to the outcomes of interest predict heaping in the running variable. Further, the estimates will be biased regardless of whether the heaps are close to the treatment threshold or (within the bandwidth but) far away.

While composition bias is not a new concern for RD designs, the type of composition bias that researchers tend to test for is of a very special type. In particular, the convention is to test for mean shifts in characteristics taking place at the treatment threshold. This diagnostic is often motivated as a test for whether or not certain types are given special treatment or better able manipulate the system in order to obtain favorable treatment. In this paper, we suggest that researchers also need to be concerned with abrupt compositional changes that may occur at heap points.

Because the usual RD-diagnostic tests are not well suited to identifying such problems, we propose a more rigorous approach to establishing the validity of RD designs when the distribution of the running variable has reporting heaps. While the importance of showing disaggregated mean plots is well established as a way to visually confirm that estimates are not driven by misspecification (Cook and Campbell 1979), we have shown multiple examples in which it pays to highlight data at reporting heaps in order to visually inspect whether

[29]For visual clarity, the sample is restricted to those working 1,000 to 2,000 hours, which roughly corresponds to those working 20-40 hours throughout the year.

they are outliers that might bias estimated effects. In addition, non-random heaping may be at work when estimated effects are sensitive to the choice of bandwidth. As a more formal diagnostic to be used when it is not obvious one way or the other, we suggest that researchers estimate the extent to which characteristics at heap points "jump" off of the trend predicted by non-heaped data.

We highlight several straightforward approaches that one might consider using to address the problem of non-random heaping once it has been diagnosed. The most robust alternative is a donut-RD approach that restricts the sample in a manner that balances covariates across the threshold by dropping those at reporting heaps. One can also focus estimation solely on the heaped data but this approach will limit the extent to which one can shrink the bandwidth. If it is necessary to pool the heaped data and continuous data for estimation, it is better to allow flexible trends and intercepts for the heaped and non-heaped data than simply controlling for heaped data with an indicator variable.

References

Aiken, L. S., S. G. West, D. E. Schwalm, J. Carroll, and S. Hsuing (1998): "Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program," *Evaluation Review*, 22(4), 207 – 244.

Almond, D., J. J. Doyle, Jr., A. E. Kowalski, and H. Williams (2010): "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns," *Quarterly Journal of Economics*, 125(2), 591–634.

——— (Forthcoming): "The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: A Reply to Barreca, Guldi, Lindo, and Waddell," *Quarterly Journal of Economics*.

Barecca, A. I., M. Guldi, J. M. Lindo, and G. R. Waddell (forthcoming): "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *The Quarterly Journal of Economics*.

Berk, R., G. Barnes, L. Ahlman, and E. Kurtz (2010): "When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment," *Journal of Experimental Criminology*, 6, 191208.

Black, D., J. Galdo, and J. Smith (2005): "Evaluating the regression discontinuity design using experimental data," *Mimeo*.

Buddelmeyer, H., and E. Skoufias (2004): "An evaluation of the performance of regression discontinuity design on PROGRESA," *World Bank Policy Research Working Paper No. 3386*.

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.

Cook, T. D., and D. T. Campbell (1979): *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Cook, T. D., and V. C. Wong (2008): "Empirical Tests of the Validity of the Regression Discontinuity Design," *Annals of Economics and Statistics*, pp. 127–150.

Dickert-Conlin, S., and T. Elder (forthcoming): "Suburban Legend: School Cutoff Dates and the Timing of Births," *Economics of Education Review*.

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

Imbens, G. W., and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142(2), 615–635.

LaLonde, R. (1986): "Evaluating the econometric evaluations of training with experimental data," *The American Economic Review*, 76(4), 604–620.

Lee, D. S., and D. Card (2008): "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 127(2), 655–674.

Lee, D. S., and T. Lemieux (forthcoming): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*.

McCrary, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142(2), 698–714.

McCrary, J., and H. Royer (2011): "The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth," *American Economic Review*, 101(1), 158–195.

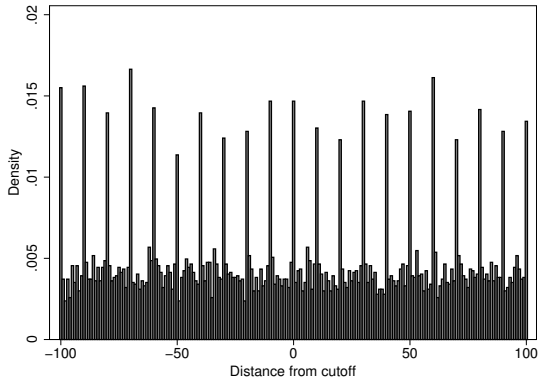
Porter, J. (2003): "Estimation in the Regression Discontinuity Model," *Mimeo*.

Shadish, W., R. Galindo, V. Wong, P. Steiner, and T. Cook (2011): "A randomized experiment comparing random to cutoff-based assignment," *Psychological Methods*, 16(2), 179–219.

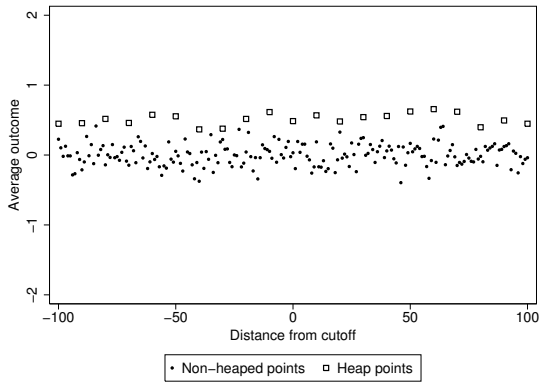
Van der Klaauw, W. (2008): "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour: Review of Labour Economics and Industrial Relations*, 22(2), 219–245.

Figure 1
Baseline Data-Generating Process

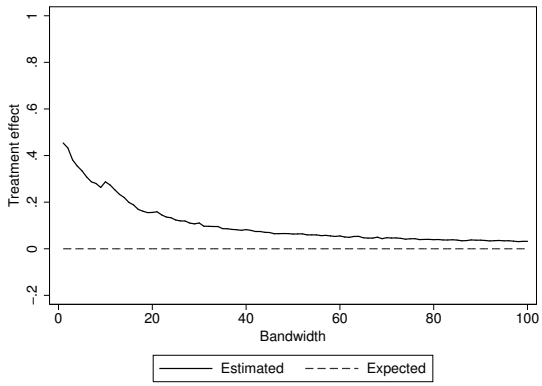
Panel A: Distribution



Panel B: Mean Outcomes



Panel C: Estimated Treatment Effect by Bandwidth

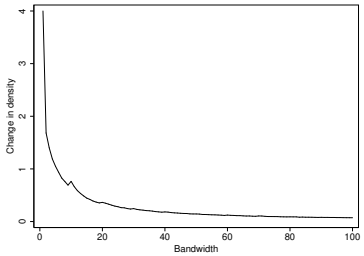


Note: 80 percent (continuous types) have R_i randomly drawn from U(-100,100), a treatment effect of zero, an expected outcome of zero, and an error term drawn from $N(0,1)$. The remaining 20 percent (heaped types) have R_i drawn from $\{-100,-90,\ldots,100\}$, a treatment effect of zero, an expected outcome of 0.5, and an error term drawn from $N(0,1)$. Panels A and B are based on a single replication with 10,000 observations. The RD-based estimates of the treatment effect (using Equation 1) in Panel C are based on 1,000 replications of randomly drawn samples of 10,000 observations.

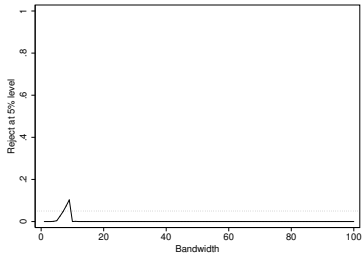
Figure 2
Standard Diagnostic Checks

Panel A: Testing For A Discontinuity in the Distribution

Estimates

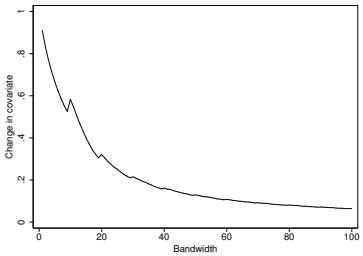


Rejection Rate at 5%

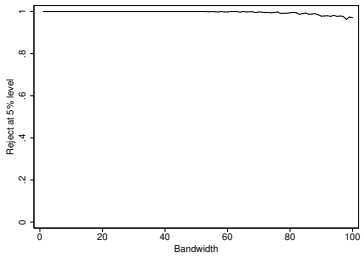


Panel B: Testing Whether Type is Balanced Across Threshold

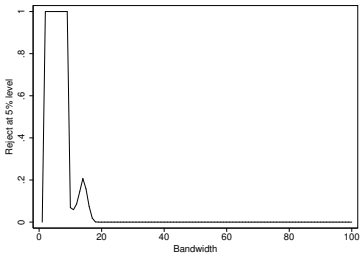
Estimates



Rejection Rate at 5% (Assuming iid Errors)



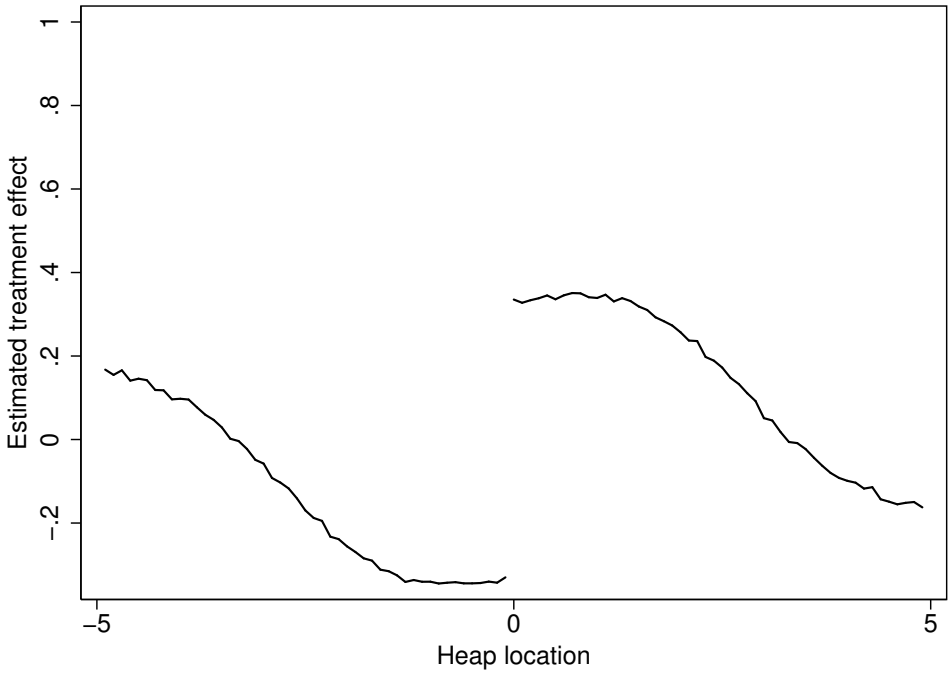
Rejection Rate at 5% (Clustering SEs)



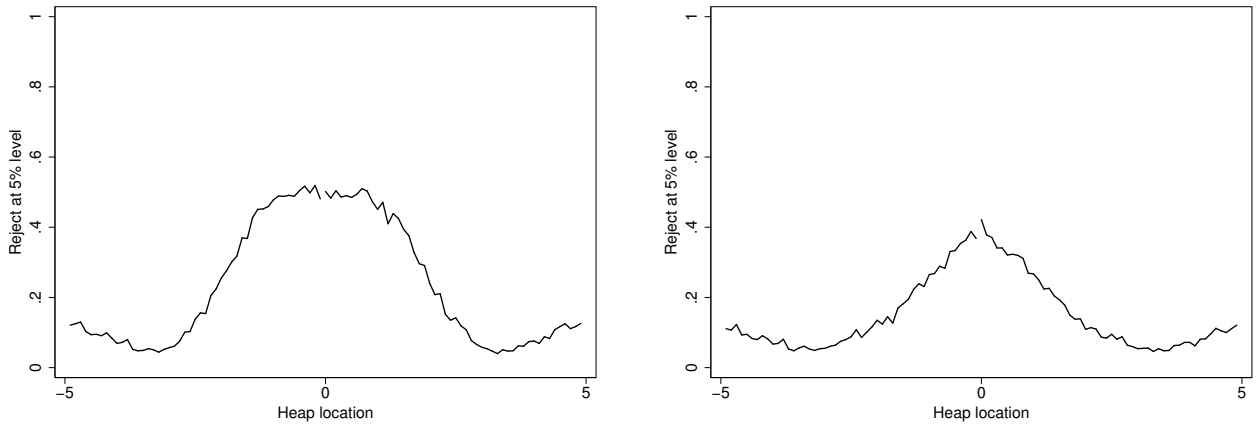
Note: The data is the same as that described in Table 1. In Panel A, data are grouped into one-unit bins to obtain frequency counts as the left-hand-side variable to be considered using estimating equation 1. In Panel, the left-hand-side variable is an indicator variable for being a "heaped type."

Figure 3
Are Estimates Biased when the Heap Does Not Fall at the Cutoff?

Panel A: Estimated Effect by Location of Heap Relative to Cutoff



Panel B: Rejection Rates at 5% Level



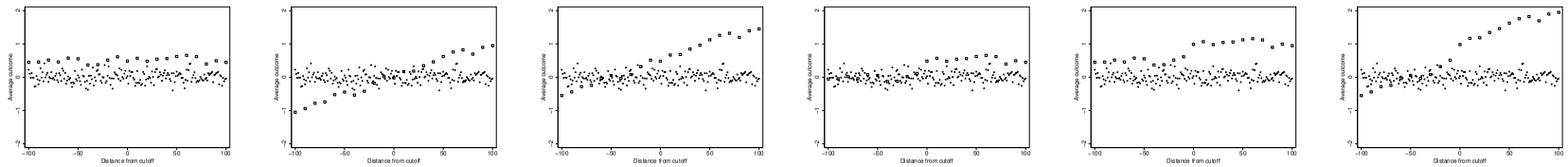
Note: The data and estimates are similar to that described in Table 1 except we consider moving the data heap nearest to treatment threshold left and right. We use bandwidths less than five to focus on the influence of a single heap.

Figure 4
Treatment Effect Estimates Based on Alternative Data-Generating Processes and Empirical Approaches

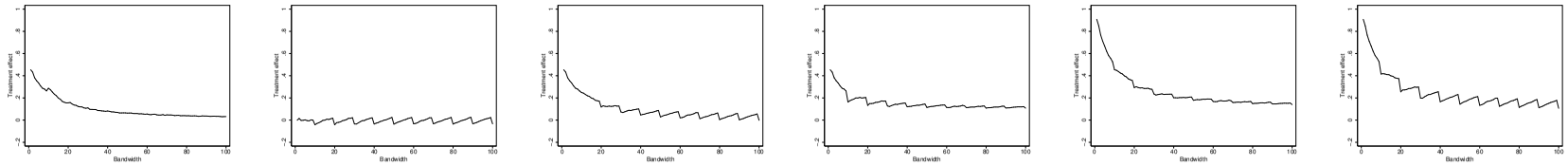
Panel A: Details of DGPs

DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-6
Average Effect = 0	Average Effect = 0	Average Effect = 0	.1	Average Effect = .1	Average Effect = .1
Effect for heaped = 0	Effect for heaped = 0	Effect for heaped = 0	Effect for heaped = .5	Effect for heaped = .5	Effect for heaped = .5
Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0
Intercept for heaped = .5	Intercept for heaped = 0	Intercept for heaped = .5	Intercept for heaped = 0	Intercept for heaped = .5	Intercept for heaped = .5
Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0
Slope for heaped = 0	Slope for heaped = .01	Slope for heaped = .01	Slope for heaped = 0	Slope for heaped = 0	Slope for heaped = .01
Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0

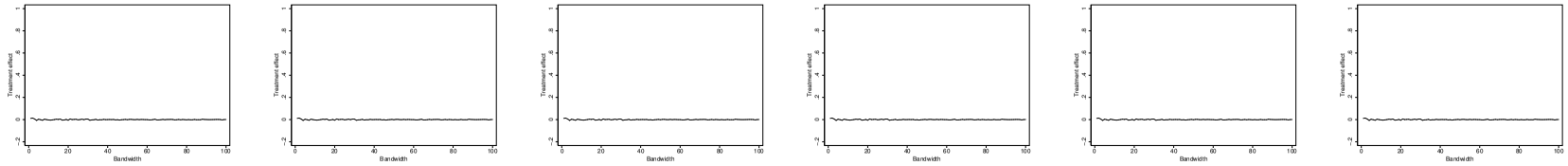
Panel B: Means



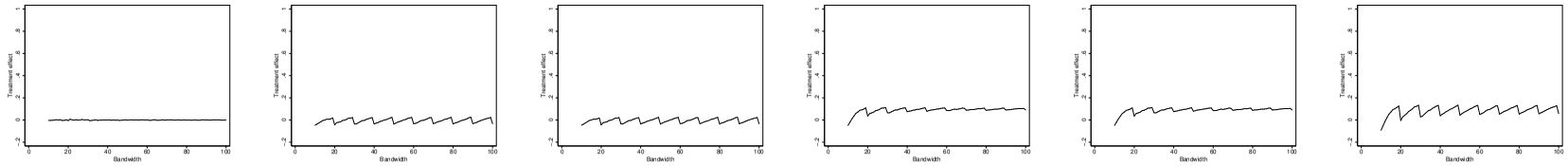
Panel C: Standard RD Estimates



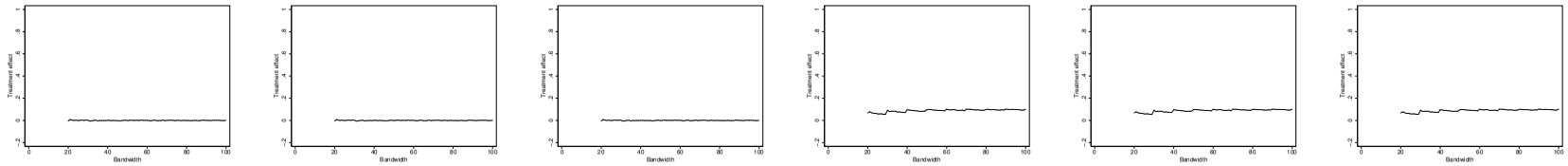
Panel D: Donut-RD Estimates Dropping Data at Heaps



Panel E: Allowing Separate Intercept for Heaped Data

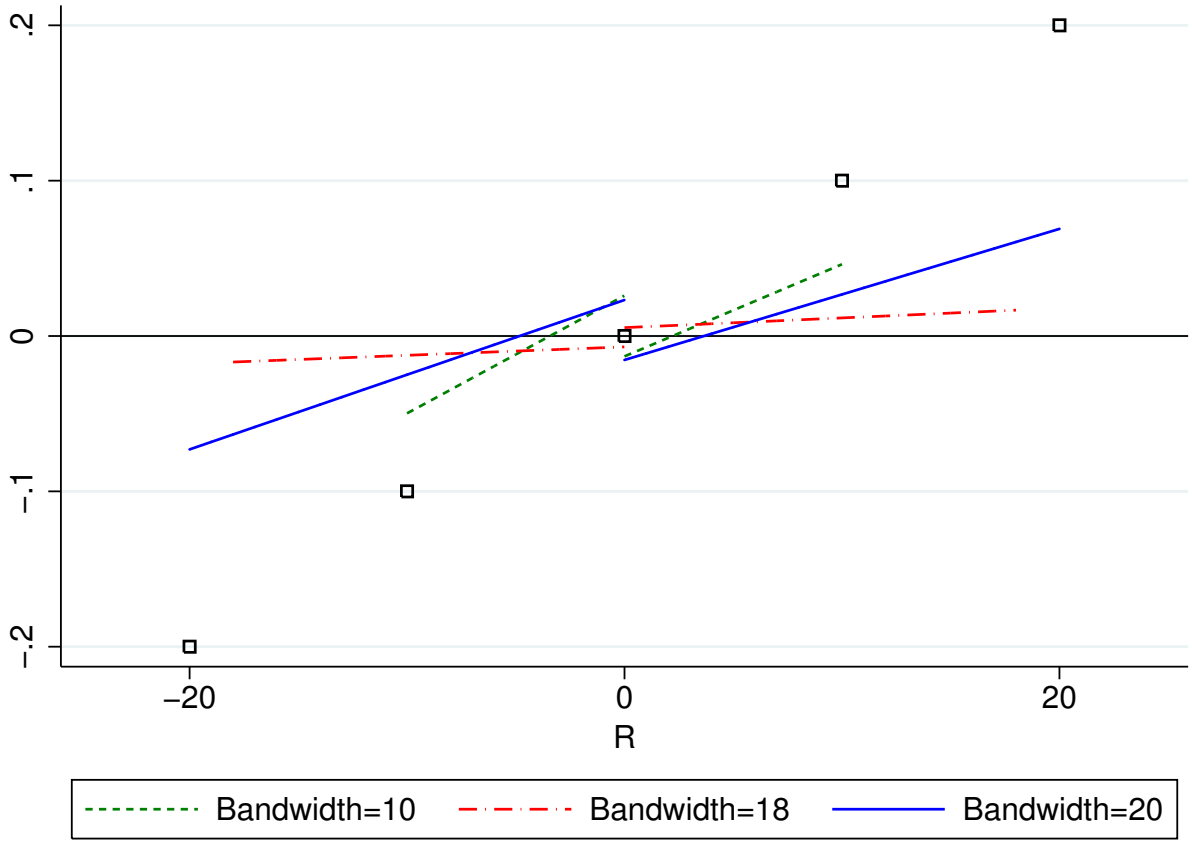


Panel F: Allowing Separate Intercept and Trends for Heaped Data



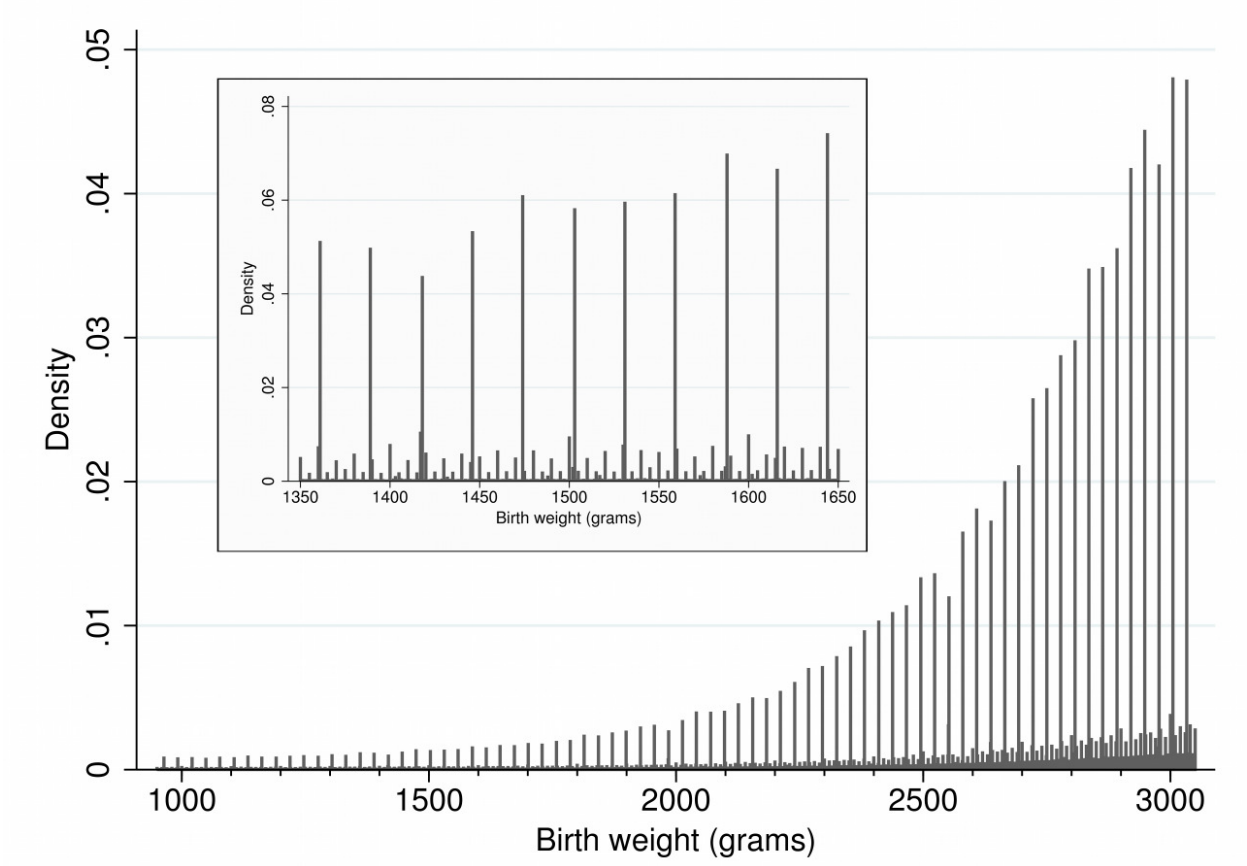
Note: All DGPs involve 80 percent (continuous types) with R_i randomly drawn from U(-100,100) and 20 percent (heaped types) with R_i drawn from $\{-100, -90, \ldots, 100\}$. Panels A and B are based on a single replication with 10,000 observations. Treatment effect estimates in panels C through F are based on 1,000 replications of randomly drawn samples of 10,000 observations.

Figure 5
Why Do DGP-2 Estimates Exhibit a Sawtooth Pattern?
Regression Lines Using Selected Bandwidths and a Standard-RD Design



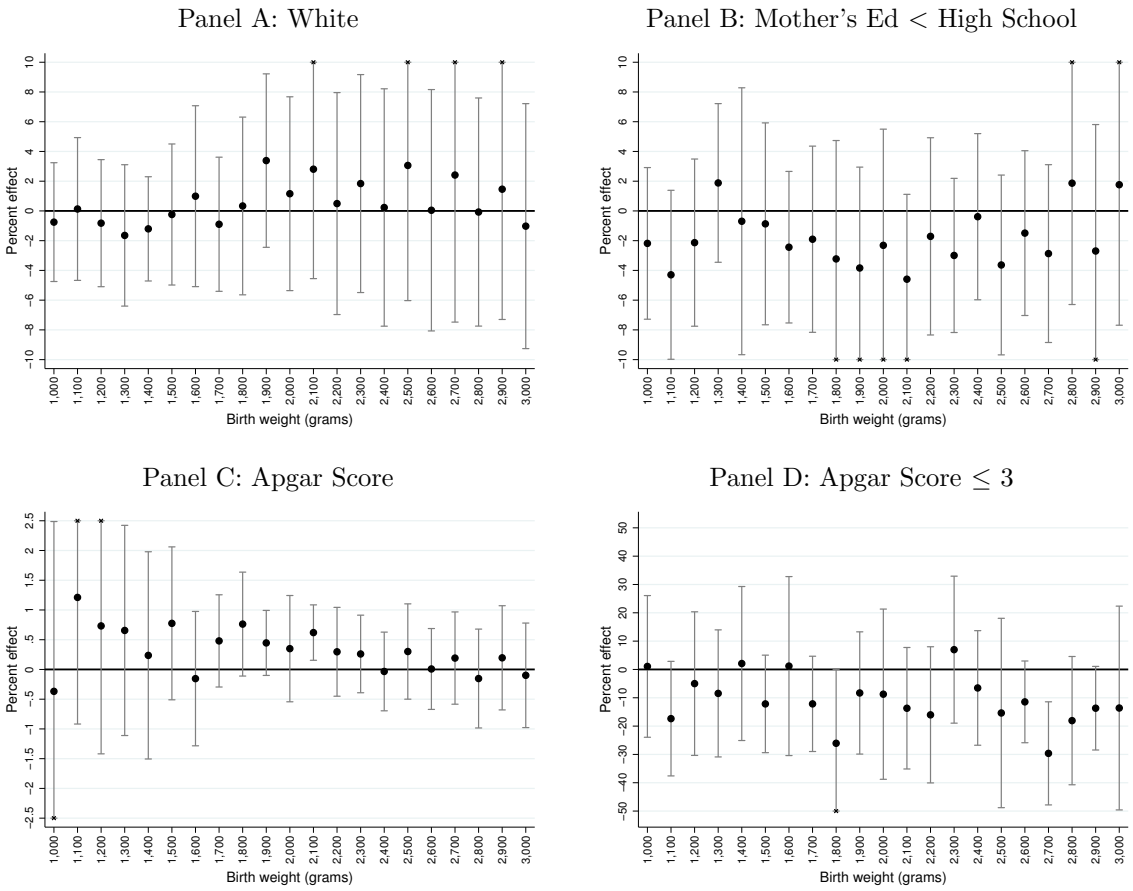
Note: Estimated standard-RD regression equations for DGP-2 are based on 1,000 replications of randomly drawn samples of 10,000 observations. Squares denote expected values at heap points.

Figure 6
Distribution of Birth Weights, Used in Almond et al. (2010)



Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

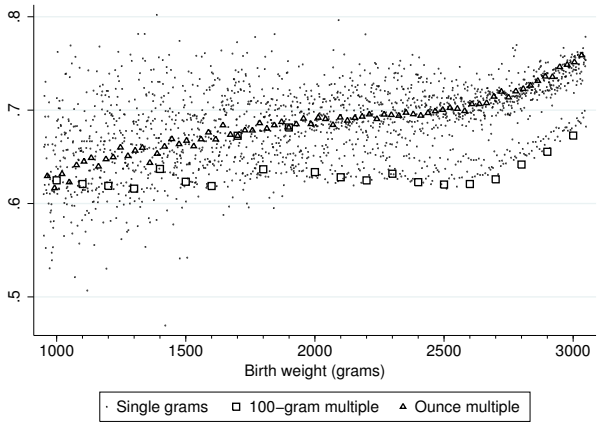
Figure 7
Standard Tests for Discontinuities in Characteristics Across Various Birth Weight Cutoffs



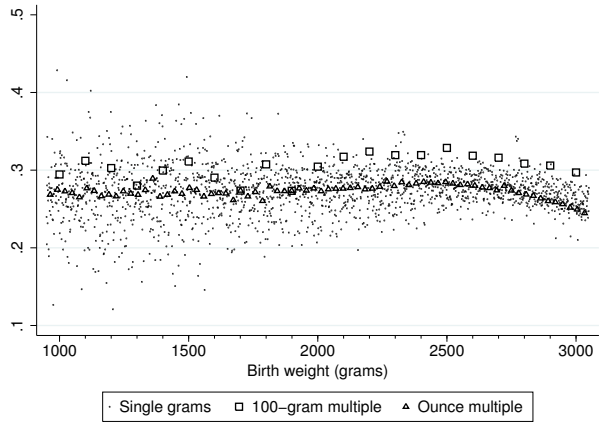
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). Following ADKW, estimates use a bandwidth of 85 grams and rectangular kernel weights, standard errors are clustered at the gram-level, and all models include a linear trend in birth weights that is flexible on either side of the cutoff.

Figure 8
Child Characteristics By Birth Weight

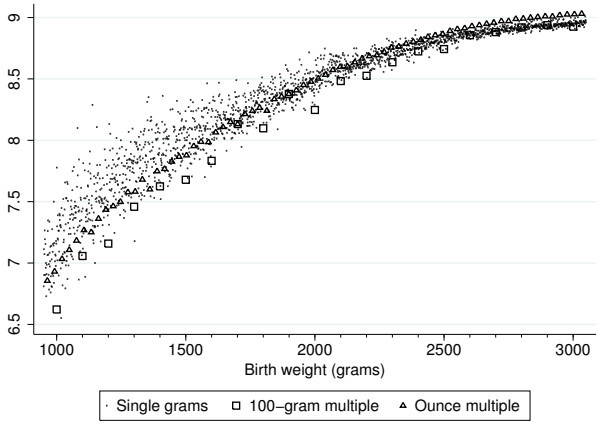
Panel A: Fraction White



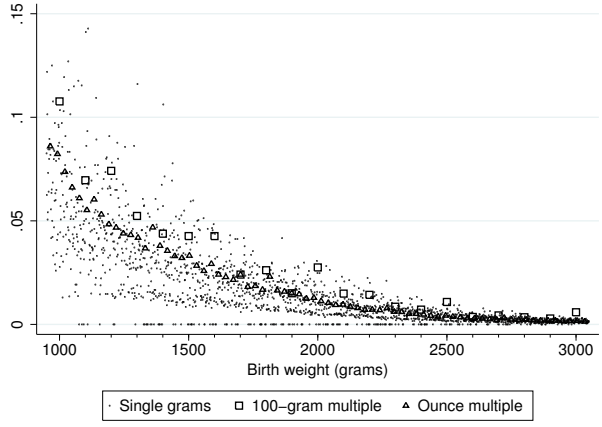
Panel B: Fraction Ed < High School



Panel C: Mean Apgar Score



Panel D: Fraction Apgar ≤ 3

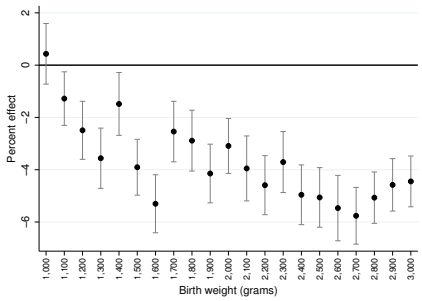


Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

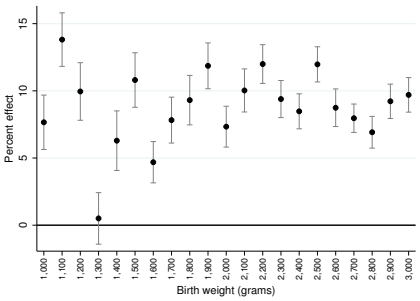
Figure 9

Panel A: Estimated Jumps in Child Characteristics at 100-Gram Birth Weight Multiples

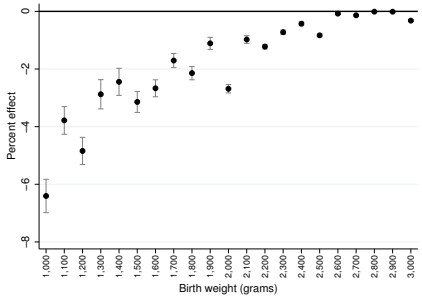
White



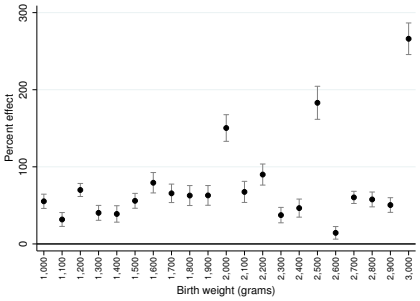
Mother's Ed < High School



Apgar Score

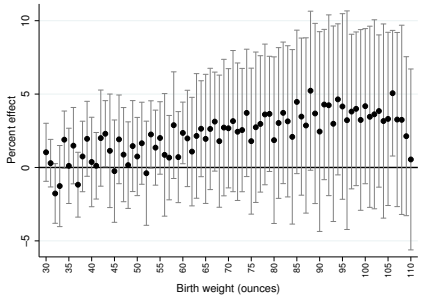


Apgar Score ≤ 3

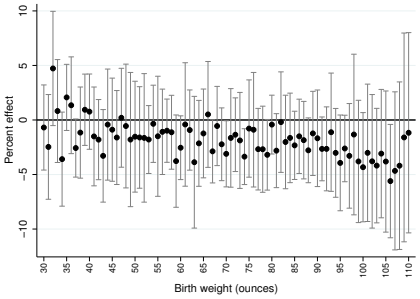


Panel B: Estimated Jumps in Child Characteristics at Ounce Multiples

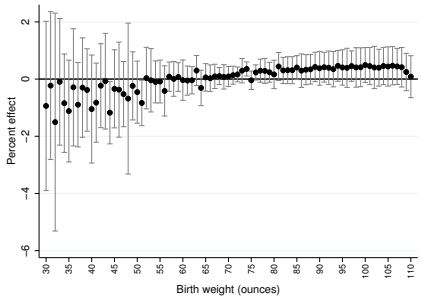
White



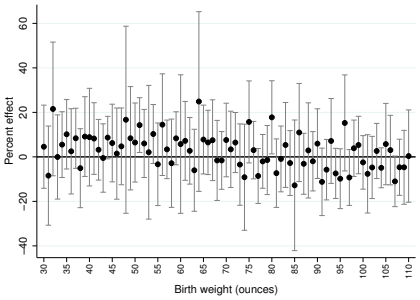
Mother's Ed < High School



Apgar Score



Apgar Score ≤ 3

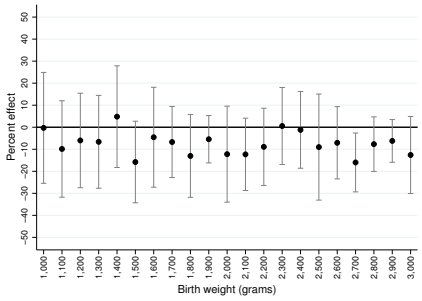


Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). Estimated "jumps" are relative to the trend which is based on a bandwidth of 85 grams. In Panel A, those at ounce multiples are excluded. Likewise, in Panel B, those at 100-gram multiples are excluded in addition to those at ounce heaps other than the one under consideration.

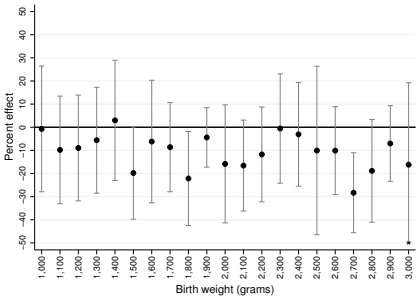
Figure 10
Estimated Impacts of Having Birth Weight < Various Cutoffs On Infant Mortality

Panel A: Standard-RD Estimates

One-Year Mortality

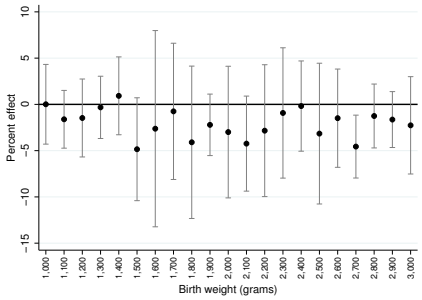


28-Day Mortality

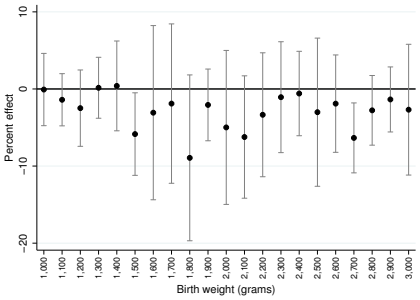


Panel B: Covariate-adjusted Estimates

One-Year Mortality

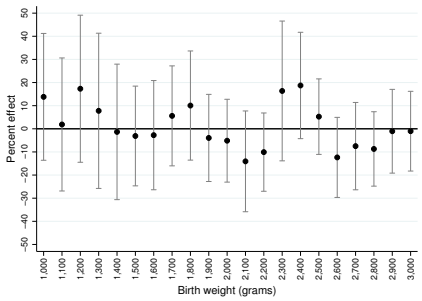


28-Day Mortality

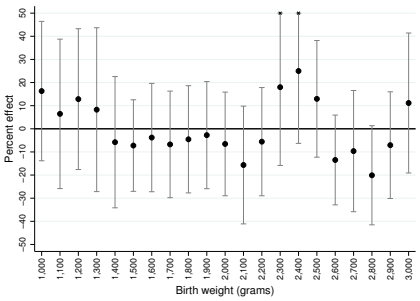


Panel C: Donut-RD Estimates

One-Year Mortality

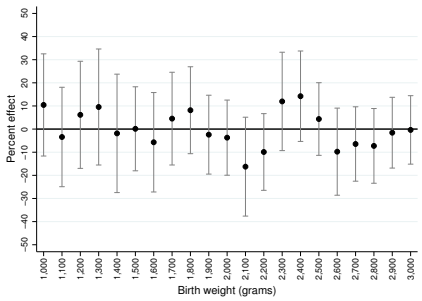


28-Day Mortality

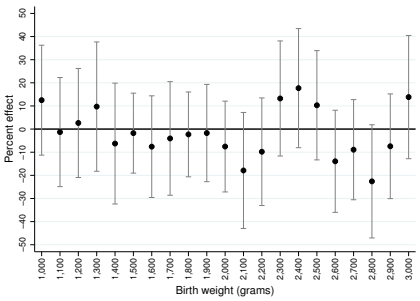


Panel D: Estimates Allowing Separate Intercepts and Trends for Those at Heaps

One-Year Mortality



28-Day Mortality

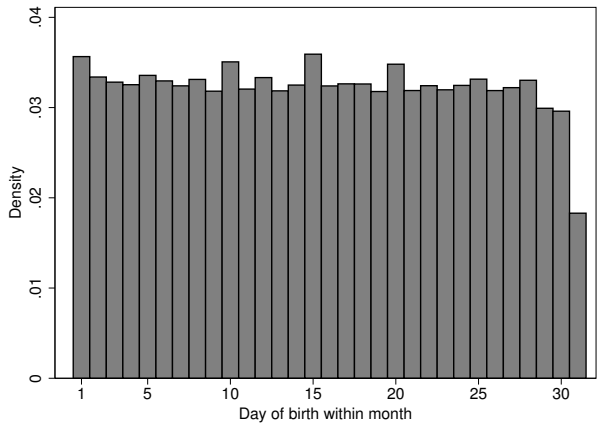


Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). Following ADKW, estimates use a bandwidth of 85 grams and rectangular kernel weights, standard errors are clustered at the gram-level, and all models include a linear trend in birth weights that is flexible on either side of the cutoff. Controls included in Panel B are described in the text. In Panel C, we omit observations at 100-gram and ounce heaps from the analysis. In Panel D, we omit from the analysis those at 100-gram multiples and allow for separate trends (and a separate intercept) for those at ounce multiples.

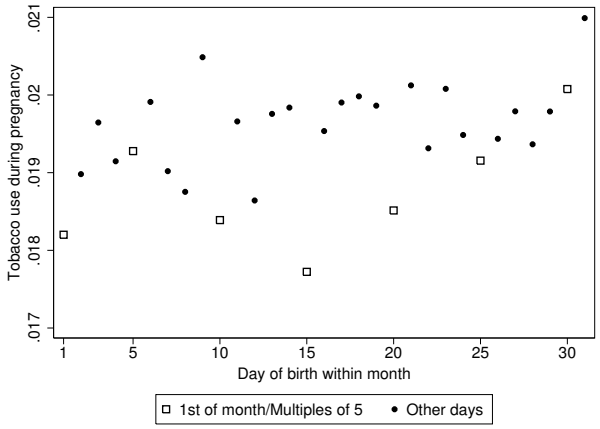
Figure 11
Non-Random Heaping In Other Data

Panel A: Mother's Date of Birth in California Birth Records, Used in McCrary and Royer (2011)

Distribution

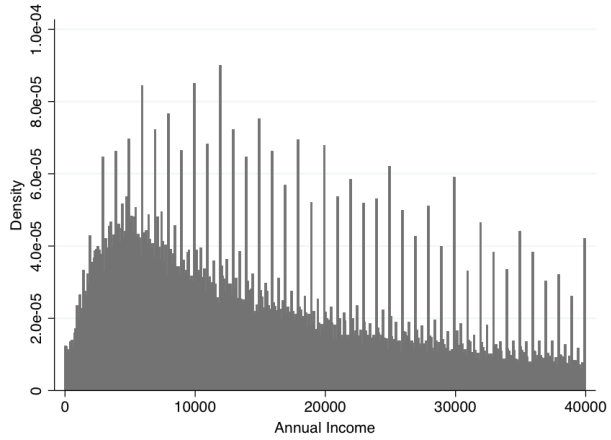


Fraction Using Tobacco During Pregnancy



Panel B: Nominal Income in PSID

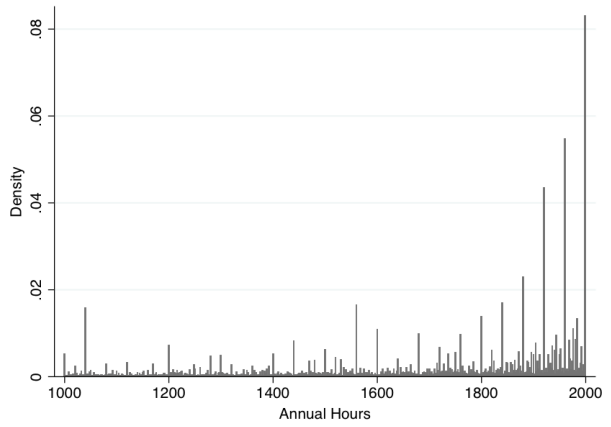
Distribution



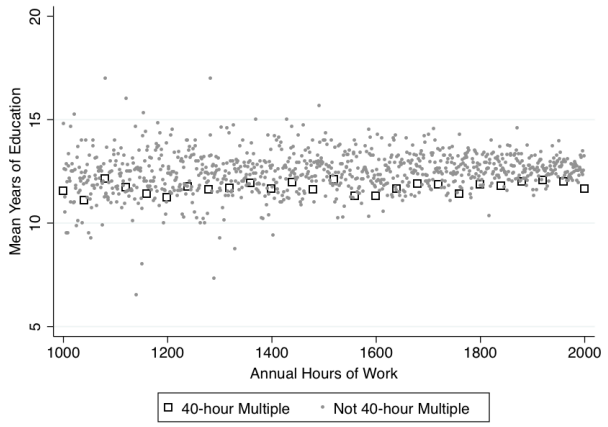
Fraction White

Panel B: Hours Worked in PSID

Distribution



Mean Completed Education



Note: Panel A uses one-day bins, Panel B uses one-hour bins, and Panel C uses $100 bins. See text for details on the data.

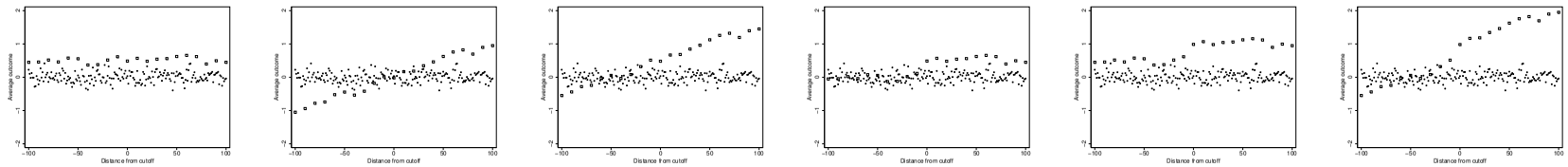
Appendix

Figure A1
Rejection Rates at 5% Level (Assuming iid Errors) Using Alternative Data-Generating Processes

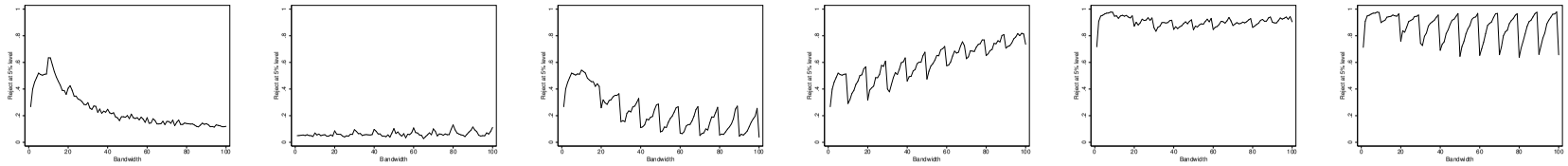
Panel A: Details of DGPs

DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-6
Average Effect = 0	Average Effect = 0	Average Effect = 0	.1	Average Effect = .1	Average Effect = .1
Effect for heaped = 0	Effect for heaped = 0	Effect for heaped = 0	Effect for heaped = .5	Effect for heaped = .5	Effect for heaped = .5
Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0
Intercept for heaped = .5	Intercept for heaped = 0	Intercept for heaped = .5	Intercept for heaped = 0	Intercept for heaped = .5	Intercept for heaped = .5
Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0
Slope for heaped = 0	Slope for heaped = .01	Slope for heaped = .01	Slope for heaped = 0	Slope for heaped = 0	Slope for heaped = .01
Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0

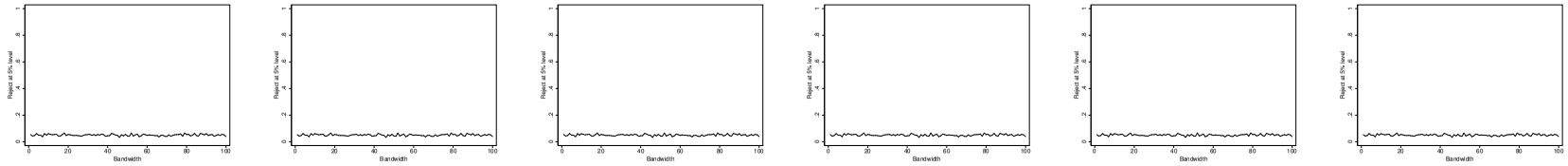
Panel B: Means



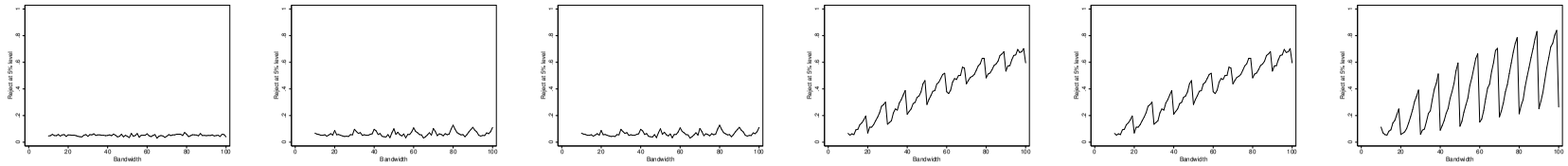
Panel C: Standard RD Estimates



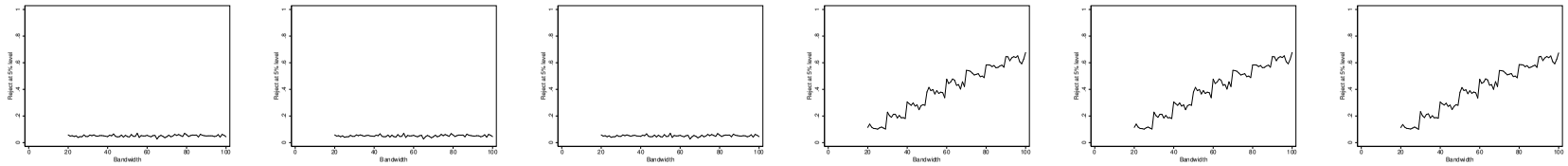
Panel D: Donut-RD Estimates Dropping Data at Heaps



Panel E: Allowing Separate Intercept for Heaped Data



Panel F: Allowing Separate Intercept and Trends for Heaped Data



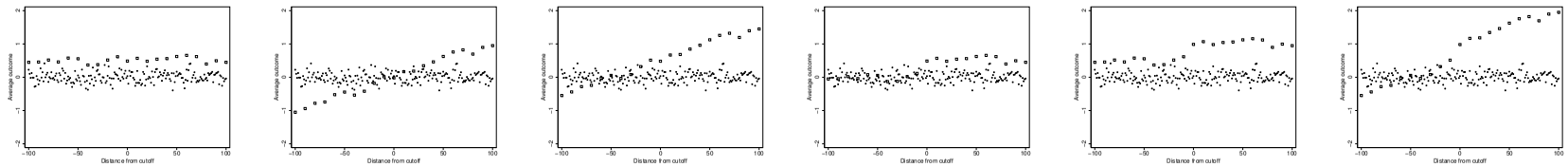
Note: All DGPs involve 80 percent (continuous types) with R_i randomly drawn from U(-100,100) and 20 percent (heaped types) with R_i drawn from $\{-100, -90, \ldots, 100\}$. Panels A and B are based on a single replication with 10,000 observations. Treatment effect estimates in panels C through F are based on 1,000 replications of randomly drawn samples of 10,000 observations.

Figure A2
Rejection Rate at 5% Level (Clustering SEs) Using Alternative Data-Generating Processes

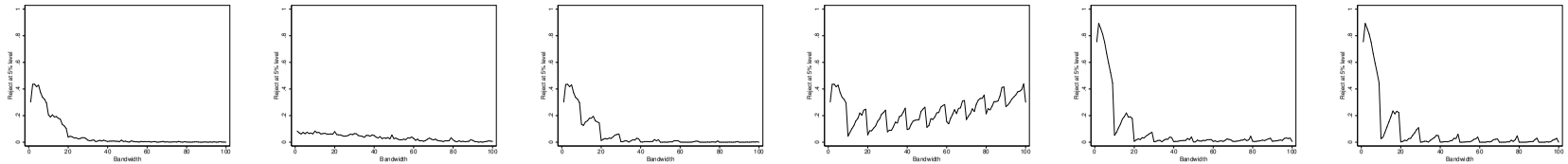
Panel A: Details of DGPs

DGP-1	DGP-2	DGP-3	DGP-4	DGP-5	DGP-6
Average Effect = 0	Average Effect = 0	Average Effect = 0	.1	Average Effect = .1	Average Effect = .1
Effect for heaped = 0	Effect for heaped = 0	Effect for heaped = 0	Effect for heaped = .5	Effect for heaped = .5	Effect for heaped = .5
Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0	Effect for cts. = 0
Intercept for heaped = .5	Intercept for heaped = 0	Intercept for heaped = .5	Intercept for heaped = 0	Intercept for heaped = .5	Intercept for heaped = .5
Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0	Intercept for cts. = 0
Slope for heaped = 0	Slope for heaped = .01	Slope for heaped = .01	Slope for heaped = 0	Slope for heaped = 0	Slope for heaped = .01
Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0	Slope for cts. = 0

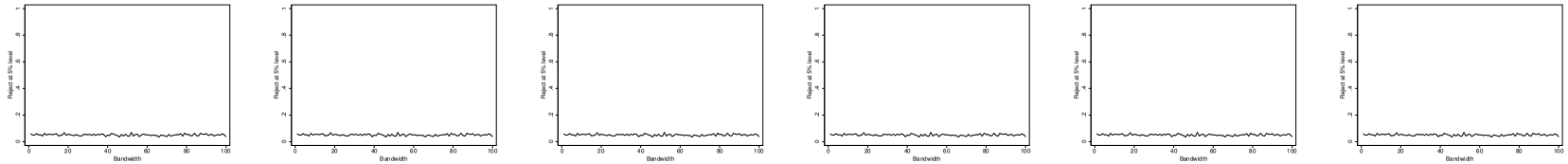
Panel B: Means



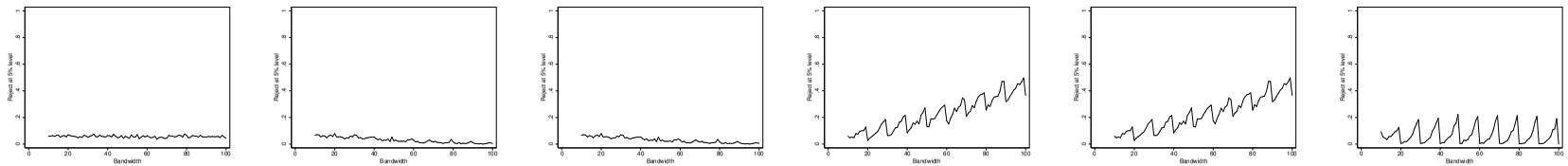
Panel C: Standard RD Estimates



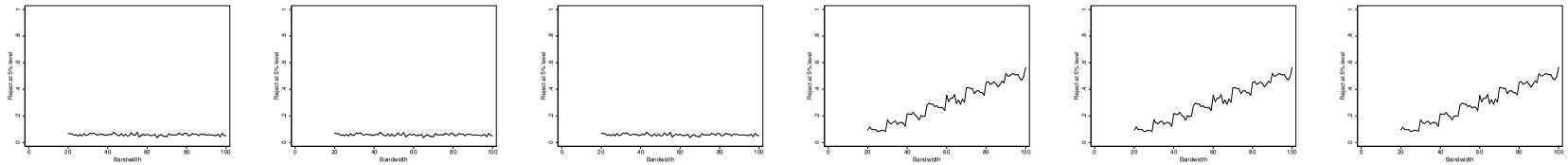
Panel D: Donut-RD Estimates Dropping Data at Heaps



Panel E: Allowing Separate Intercept for Heaped Data



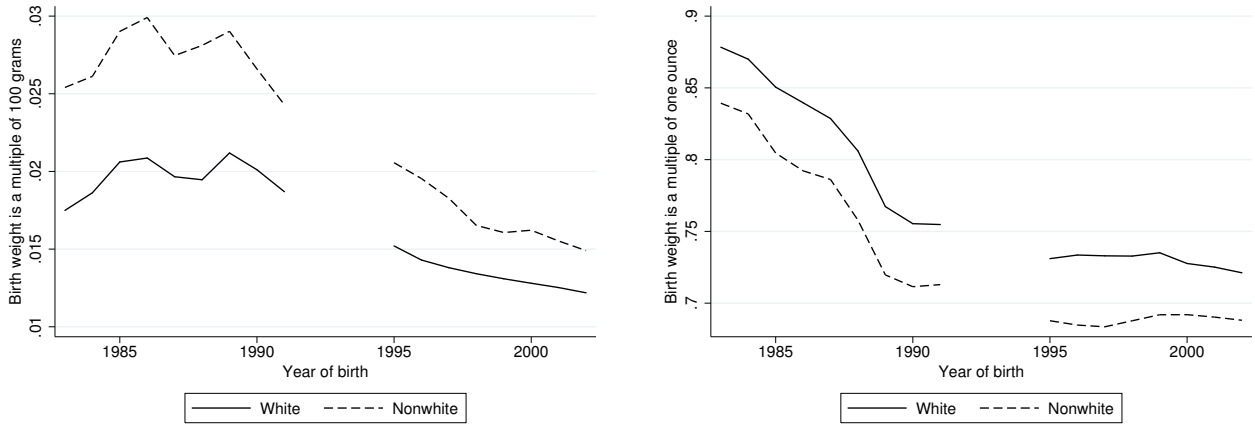
Panel F: Allowing Separate Intercept and Trends for Heaped Data



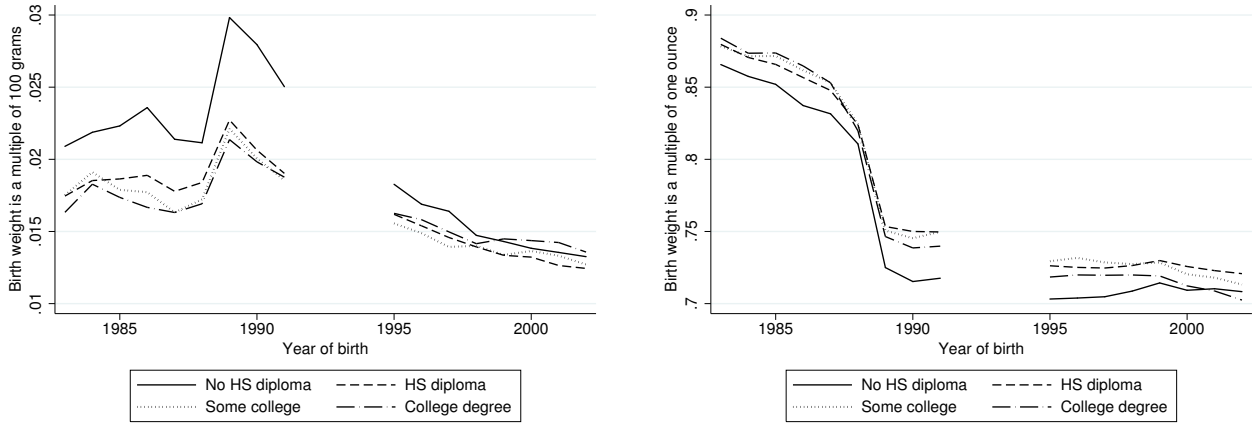
Note: All DGPs involve 80 percent (continuous types) with R_i randomly drawn from U(-100,100) and 20 percent (heaped types) with R_i drawn from $\{-100, -90, \ldots, 100\}$. Panels A and B are based on a single replication with 10,000 observations. Treatment effect estimates in panels C through F are based on 1,000 replications of randomly drawn samples of 10,000 observations.

Figure A3
Fraction of Births Recorded in 100s of Grams and Ounces Over Time

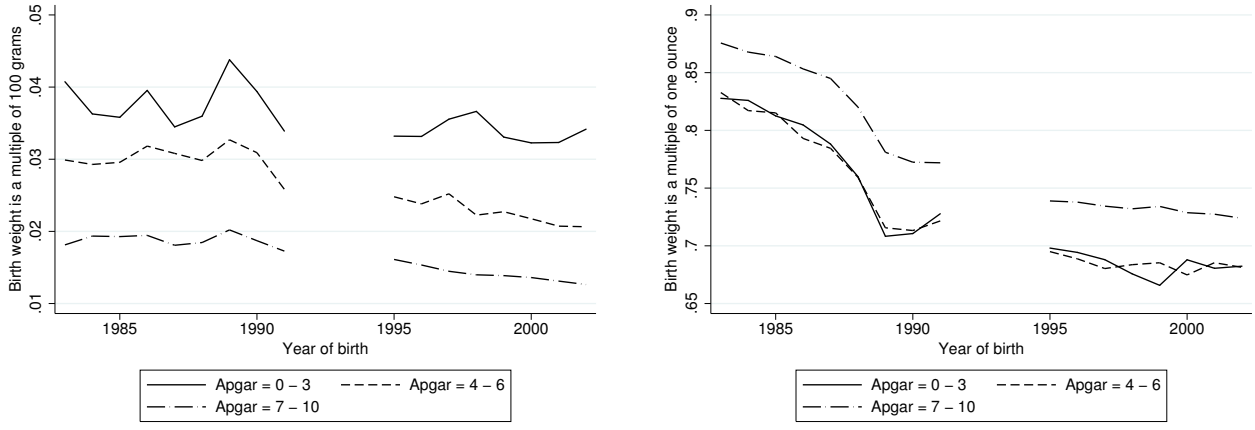
Panel A: By Race



Panel B: By Mother's Education



Panel C: By Apgar Score



Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).