DISCUSSION PAPER SERIES

IZA DP No. 5106

Running *and Jumping* Variables in RD Designs: Evidence Based on Race, Socioeconomic Status, and Birth Weights

Alan Barreca
Melanie Guldi
Jason M. Lindo
Glen R. Waddell

August 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Running *and Jumping* Variables in RD Designs: Evidence Based on Race, Socioeconomic Status, and Birth Weights

Alan Barreca
Tulane University

Melanie Guldi
Mount Holyoke College

Jason M. Lindo
University of Oregon
and IZA

Glen R. Waddell
University of Oregon
and IZA

Discussion Paper No. 5106
August 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5106
August 2010

ABSTRACT

Running *and Jumping* Variables in RD Designs: Evidence Based on Race, Socioeconomic Status, and Birth Weights*

Throughout the years spanned by the U.S. Vital Statistics Linked Birth and Infant Death Data (1983-2002), birth weights are measured most precisely for children of white and highly educated mothers. As a result, less healthy children, who are more likely to be of low socioeconomic status, are disproportionately represented at multiples of round numbers. This has crucial implications for any study using a regression discontinuity design in which birth weights are used as the running variable. For example, estimates will be biased in a manner that leads one to conclude that it is "good" to be strictly to the left of any 100-gram cutoff. As such, prior estimates of the effects of very low birth weight classification (Almond, Doyle, Kowalski, and Williams 2010) have been overstated and appear to be zero. This analysis highlights a more general problem that can afflict regression discontinuity designs. In cases where attributes related to the outcomes of interest predict heaping in the running variable, estimated effects are likely to be biased. We discuss approaches to diagnosing and correcting for this type of problem.

JEL Classification: C21, C14, I12

Keywords: regression discontinuity, donut RD, birth weight, infant mortality

Corresponding author:

Glen R. Waddell
Department of Economics
University of Oregon
Eugene, OR 97403-1285
USA
E-mail: waddell@uoregon.edu

* The authors thank Doug Almond, Joe Doyle, Todd Elder, Hilary Hoynes, Amanda Kowalski, Thomas Lemieux, Justin McCrary, Doug Miller, Marianne Page, Larry Singell, Ann Huff Stevens, and Heidi Williams for their comments and suggestions.

1 Introduction

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., incomes, ages, heights), when tools with limited precision are used for measurement (e.g., birth weights, pollution, rainfall), and when continuous data are rounded or otherwise discretized (e.g., letter grades, grade point averages). We also see heaping in many idiosyncratic variables, such as day of birth where there are relatively few births on weekends and holidays when medical procedures are rarely scheduled, work activity where there tends to be heaping at 40 hours per week after which employers are required to pay overtime wages, and wages where heaping occurs at state-set minimum wages. While ignoring heaping may be innocuous in many circumstances, in this paper we show that doing so can have serious consequences. In particular, in regression discontinuity (RD) designs, estimates are likely to be biased if attributes related to the outcomes of interest predict heaping in the running variable.

As an example, we re-evaluate the effects of a child being classified as very low birth weight (i.e., having measured birth weight strictly less than 1500 grams) on infant mortality. As explained in Almond, Doyle, Kowalski, and Williams (2010), hereafter ADKW, this is a topic of great importance because hospitals use very low birth weight status to determine treatment intensity, either through hospital protocol or as a rule of thumb. While ADKW's RD estimates suggest that very low birth weight classification reduces infant mortality, our results indicate that this finding is driven by composition bias related to systematic heterogeneity in birth weight measurement.

Focusing on the measurement of birth weights in the United States, we show that heaping at round numbers has been a persistent feature of the data since at least 1983. Further, we show that birth weights are measured most precisely for children of white and highly educated mothers. As a result, low socioeconomic status children who tend to be less healthy

are disproportionately represented at 100-gram and one-ounce multiples. This non-random heaping introduces composition bias to local RD estimates.

In demonstrating this bias, we estimate the effects of having birth weights strictly less than each 100-gram cutoff between 1000 and 3000 grams. Because of the non-random heaping at 100-gram multiples, nearly all such estimates suggest that children with birth weights below the cutoffs have more-favorable mortality outcomes. The estimated effect at the very low birth weight cutoff does not stand out among the rest and turns out to also be biased by non-random heaping at ounce multiples as the 53-ounce heap also falls immediately to the right of the cutoff (at 1503 grams). In the end, we show that very low birth weight classification has no impact on infant mortality when estimates use variation that does not exhibit composition bias.

We also show that standard approaches to verifying the validity of the RD research design may be insufficient for diagnosing this type of problem. For this reason, when there are heaps in the distribution of the running variable, we suggest that researchers verify that such heaps are not related to characteristics that might predict the outcomes of interest. In addition, we argue that researchers should examine whether there are discontinuities where they are not expected – at placebo cutoffs, both for outcomes of interest and for covariates. Finally, we show that a "donut RD design" effectively restricts the sample in a way that reduces the composition bias introduced by this type of non-random sorting.

The rest of the paper is organized as follows. In Section 2, we describe the precision with which birth weights are measured, how this relates to socioeconomic status, and how the relationship has changed over time. In Section 3, we demonstrate the resulting bias in a RD design in which birth weights are the running variable. In Section 4, we discuss the failure of "conventional RD validity checks" to diagnose this type of problem. In Section 5, we briefly provide two additional examples of situations in which similar problems could arise. We return to the general problem and offer concluding remarks in Section 6, highlighting

methods that can be used to diagnose and correct for the potential bias.

2 The Measurement of Birth Weights

Birth weights are typically measured using a hanging scale, a balance scale, or a digital scale. Scales are 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. Figure 1 shows the distribution of birth weights using data from the Vital Statistics of the United States from 1983–2002.[1] This figure shows heaping at 100-gram and ounce multiples, with the latter being most dramatic.

As one would expect, scale prices are strongly related to their resolutions. Today, the least-expensive scales cost under one hundred dollars while the most expensive cost approximately two thousand dollars. For this reason, we would expect more-precise estimates at hospitals with greater resources, or hospitals that tend to serve more-affluent patients.[2]

Panels A and B of Figure 2 provide extensive evidence that the precision with which birth weight is measured is strongly related to socioeconomic status. Specifically, these figures plot the fraction of children with birth weights reported in exact multiples of 100 grams over time, with Panel A stratifying on race and Panel B stratifying on mother's education. This figure shows that birth weights tend to be measured more precisely for whites than nonwhites.[3]

[1]Note that data is not available for 1992–1994. For an in depth description of the data, see ADKW.

[2]With general improvement in technology one would anticipate that measurement would appear more precise in the aggregate over time. We have verified that this is indeed the case.

[3]Note that a major reason that the figure does not show smooth trends is because data is not consistently

The gap has closed over time but remains evident throughout the sample period. Similarly, the precision with which birth weights are estimated is related to mother's education. Most strikingly, birth weights have historically been measured with far less precision for children of mothers with less than a high-school education. However, this gap seems to have largely closed by the late 1990s.

Panel C demonstrates that measurement precision is also strongly related to children's health at the time of birth. This figure, with construction similar to panels A and B, stratifies on children's Apgar scores, an index of newborns' health taken immediately after delivery. In short, birth weights tend to be measured with greater precision for healthier children, or children with higher Apgar scores. With panels A and B, Panel C corroborates the well-established fact that health outcomes are closely related to socioeconomic status.[4]

As a more-rigorous way of exploring the extent to which the composition of children changes abruptly at reporting heaps, we estimate the following regression equation:

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

for $Z = 1000, 1100, ..., 3000$ where X_i is a characteristic of individual i with birth weight BW_i. Note that this regression equation is not intended to detect a mean shift across Z but, rather, the extent to which characteristics at Z differ from what would be expected based on surrounding observations. We consider observations within 85 grams of Z. We discuss the reason for choosing this specification below, in Section 3.

The results from this regression analysis confirm that child characteristics change abruptly at 100-gram multiples. Figure 3 shows estimated percent changes, γ_1/γ_0, for the probability that a mother is white, the probability that she has education less than high school, average

available for all states.

[4]Apgar scores are not available for all state-years in the Vital Statistics. However, Apgar scores are reported for the majority of births – approximately 75 percent.

Apgar score, and the probability of having an Apgar score weakly less than three out of ten.[5] For nearly every estimate, bootstrapped standard error estimates clustered at the gram level are small enough to reject that the characteristics of children at Z are on the trend line.

As further evidence that children with recorded birth weights at 100-gram and ounce multiples are systematically different, Figure 4 plots one-year mortality rates against exact birth weights. Consistent with what we would expect based on figures 2 and 3, this figure shows that children with recorded birth weights at these multiples have higher mortality rates. Similar figures for mother's race, education, and children's Apgar scores are presented in figures A1–A4 in the appendix.

3 When Non-random Measurement Precision Leads to Bias

In the previous section, we showed that there is a strong relationship between characteristics that predict infant mortality and the precision with which birth weights are measured. In this section, we show how the resulting non-random heaping in the running variable can bias RD estimates.

We estimate the effect of having a birth weight strictly less than Z on child mortality outcomes using the following regression equation:

$$Y_i = \alpha_0 + \alpha_1 1(BW_i < Z) + \alpha_2 1(BW_i < Z) * (BW_i - Z) + \alpha_3 (BW_i - Z) + \epsilon_i \qquad (2)$$

where Y_i is an outcome measure for child i with birth weight BW_i. We use a bandwidth of 85 grams and rectangular kernel weights. This follows the specification reported in ADKW

[5]To provide a frame of reference, children in this category have a 58 percent chance of surviving for at least a year while children with Apgar scores greater than 3 have a 99 percent chance of surviving for at least a year.

although, while they focus on Z=1500, we consider Z = 1000, 1100, ..., 3000.[6]

Figure 5 presents the estimated percent impacts, α_1/α_0, on one year mortality, 28-day mortality, one-week mortality, and 24-hour mortality. We bootstrap standard errors and cluster on exact grams.[7] These figures suggest that, near Z, children with birth weights less than Z routinely have better outcomes than those with birth weights (weakly) just above Z. Given that these results are largely driven by a systematically different composition of children at the cutoff, the estimated effects are much larger in magnitude when one uses a narrow bandwidth, as shown in Figure A5 in the appendix. Moreover, we note that of the 84 point estimates shown in Figure A5, 84 fall below zero.

To analyze the impact of hospital care on children's outcomes, it is reasonable for ADKW to focus on the "very low birth weight" threshold at 1500 grams given the discontinuities in treatment provision they observe around this cutoff. After estimating significant discontinuities in treatment provision at the 1500-gram threshold, ADKW then estimate discontinuities in infant mortality at the same threshold, often finding significant discontinuities which can be interpreted as evidence that the increased treatment reduces infant mortality. However, as we showed in Figure 5, analysis of almost any cutoff that is a multiple of 100 grams will suggest that children have more-favorable health outcomes when their birth weight is below a given cutoff, which we argue is due to the non-random heaping of recorded birth weights. In addition, the magnitude of the estimated effect at the 1500-gram threshold is not particularly striking when compared to the placebo thresholds, which is consistent with it being driven by underlying compositional differences.

A natural way of dealing with this composition bias is to drop observations with birth weights recorded in round numbers. While a drawback of this approach is that it cannot tell us about the effect of very low birth weight classification on the types of infants with round

[6] ADKW also present results based on local linear regressions that control for covariates. Their estimates are not very sensitive to these alternative specifications, however.

[7] Again, this is consistent with ADKW, who follow Lee and Card (2008).

birth weights, it is consistent with the usual motivation for RD designs. Specifically, we focus on what might be considered a narrow sample in order to be confident that treatment is exogenous. Figure 6 shows the estimated effects after dropping children with recorded birth weights in 100s of grams or in single ounces. While the earlier estimates (Figure 5) were negative for most of the placebo cutoffs, these estimates (Figure 6) resemble the white-noise process we would anticipate. Thus, our results indicate that the sample restrictions we employ reduce the bias produced by the non-random heaping described above.

Table 1 concentrates on the very low birth weight threshold, in particular. ADKW's estimates in Panel A and our replication of their estimates in Panel B suggest that very low birth weight classification significantly reduces mortality. The estimates in Panel C, which adds an extensive set of controls that ADKW consider, are slightly smaller but continue to indicate that low birth weight classification reduces mortality.[8]

In Panel D, we add Apgar score fixed effects. This specification is motivated by our earlier results which showed that children at ounce and 100-gram multiples have systematically worse health outcomes than others. As such, it is of utmost importance to control for underlying health conditions at birth so that estimated mortality effects are not driven by the abrupt composition changes occurring near the cutoff. We acknowledge, however, that there is a tradeoff inherent in controlling for Apgar scores. Specifically, there is a chance we are "over-controlling" since it is not impossible for 5-minute Apgar scores to be affected by treatment induced by very low birth weight classification. However, as we will discuss in the next section, Apgar scores are systematically lower to the left of almost all 100-gram cutoffs which suggests that lower Apgar scores to the left of the 1500-gram cutoff are probably not due to treatment induced by very low birth weight classification.[9] In addition, our

[8]These controls include measures of prenatal care, mother's age, mother's education, father's age, child gender, gestational age, mother's race, plurality of births, birth order, and year.

[9]See Figure 7. In addition, the means shown in figures A3 and A4 demonstrate infants at reporting heaps have systematically lower Apgar scores than those with similar but more-precisely measured birth weights. This is true when comparing means using adjacent observations with slightly higher or lower reported birth

discussions with a pediatric specialist revealed that, whereas Apgar scores are almost always taken five minutes after birth, birth weight measurements tend not to be a priority for at-risk newborns and that these children are often not weighed until well after they have been stabilized.[10] This also suggests that it is unlikely that birth weights are measured quickly enough to trigger medical intervention that would improve Apgar scores within five minutes of a child's birth.

With the caveat above, we note that the estimates controlling for Apgar scores fall dramatically and lose statistical significance.[11] If one takes Apgar scores as an exogenous measure of infant health at birth, then the fact that the estimated effects are sensitive to the inclusion of this control raises a serious concern about the RD identification strategy in this setting. The validity of the RD research design hinges crucially on the assumption that underlying characteristics related to outcomes are smooth near the cutoff determining treatment. When this assumption holds, the inclusion of covariates should only affect standard error estimates.

To the extent to which the set of controls used in Panel D capture all of the underlying characteristics related to mortality outcomes that may not vary smoothly with the running variable, the estimates should provide unbiased estimates of the effect of very low birth weight classification. However, the fact that estimates are sensitive to the inclusion of Apgar scores suggests that there might be important unobservable characteristics whose omission will bias the estimates. In other words, despite all of our controls, we still have reason to be concerned that the composition of children might not vary smoothly in birth weights.

Since we are especially concerned with composition bias involved with heaping, panels E

weights than those at reporting heaps.

[10]This certainly does not mean that low birth weight classification cannot affect infant health since it may very well be used for some hospital protocols (e.g., diagnostic ultrasounds).

[11]Note that the sample falls by approximately 20 percent moving from the estimates in Panel C to the estimates in Panel D. As such, we have verified that the change in the estimates is not driven by the change in the sample and make these results available by request.

through H either control for observations at reporting heaps with fixed effects or by dropping them from the analysis.[12] In panels E and F we address the heap at 100 grams, and the estimates are substantially smaller than the baseline estimates in panels A through C. This again suggests that the RD design is not appropriate for the full sample. Since the essence of an RD-based estimate is a comparison of mean outcomes approaching the cutoff from both sides, the estimate should not be sensitive to the observations that fall exactly at the cutoff.[13]

As we have alluded to above, the heaping at ounce intervals is also of special concern in this setting because the large heap at 53 ounces falls immediately to the right of the cutoff, at 1503 grams. For this reason, Panel G controls for reporting heaps with unrestricted fixed effects for birth weights that fall exactly at 100-gram and one-ounce multiples while Panel H drops the observations from the analysis altogether.[14] The estimated effects in these panels are extremely close to zero, revealing that very low birth weight classification has no measurable effect on infant mortality when controlling for observable and unobservable characteristics related to heaping.[15] As a whole, when we use variation that does not introduce composition bias, our estimates suggest that any treatment induced by very low birth weight classification has no impact on infant mortality.[16]

[12]We note that these two approaches are nearly equivalent and produce similar estimates.

[13]In addition, we note that only two percent of the observations fall exactly at 1500-gram cutoff.

[14]We should note that the importance of controlling for ounce multiples is not obvious and that we can not make a general statement about the direction of the bias caused by heaping at ounce multiples when considering cutoffs that are multiples of 100 grams. The bias caused by such heaping will depend crucially on where ounce multiples fall relative to the cutoff. Bias is a concern despite the fact that observations with birth weights recorded in exact ounces fall within the 85 gram bandwidth on both sides of any threshold (as one ounce corresponds to approximately 28 grams). Because the discontinuity is estimated by a linear regression allowing for different slopes on each side of the cutoff, compositional differences across the threshold will not necessarily "balance one another out" when the heaps are not symmetric on each side of the cutoff. This point is further addressed in the next section.

[15]We also note that the standard error estimates are much smaller in this panel when we cluster on one-ounce bins which is consistent with the bouncing back and forth of covariates in Figure 7 which is discussed in the next section.

[16]In an attempt estimate the treatment effect using a broader sample, we have explored restricting the sample to those at ounce multiples but this approach fails the test for balanced covariates when looking at mother's education and Apgar scores.

4 Diagnosing the Problem

These results raise the important question: why might standard RD falsification tests fail to identify the non-random sorting we describe above? For example, as suggested by Imbens and Lemieux's (2008), Angrist and Pischke's (2009), and Lee and Lemieux's (forthcoming) "Guides to Regression Discontinuity Designs," ADKW consider whether there are discontinuities in the distribution of birth weights and in observable characteristics, finding no cause for concern. So what went wrong?

ADKW themselves note that there are heaps at round-gram numbers and at gram equivalents of ounce multiples. The heaps are quite noticeable upon visual inspection of the distribution of birth weights (shown in ADKW's Figure 1 which is similar to our Figure 1). As such, one might anticipate a red flag via McCrary's (2008) estimation procedure to test for non-random sorting in RD designs by considering whether or not the distribution is discontinuous at the treatment threshold. However, ADKW find that estimated discontinuity in the distribution is not statistically significant.[17] This is likely due to the general lumpiness of the birth weight data – heaping is not only observed at 100-gram multiples and, further, the heaping at 100-gram multiples is actually quite small compared to the heaping observed at ounce multiples.

In addition to the McCrary test, 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 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

[17] They report a discontinuity estimate of -2100 with a standard error estimate of 1500.

thresholds, it makes sense to conclude that the heaping observed at the 1500-gram threshold is "normal." The problem that this example make evident, however, is that even if agents are not manipulating the running variable around these thresholds in a strategic manner, the heaping at *all* round numbers is non-random.

Panels A and B of Figure 7 replicate ADKW's analysis of covariates at the 1500-gram cutoff along with the placebo cutoffs. In particular, these figures use the same approach used to estimate the impacts on infant mortality to estimate mean shifts in child characteristics across the considered thresholds. These figures 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 education less than high school. However, the set of estimates reveal a distinct pattern that might serve as a red flag; Panel A of Figure 7, and to a lesser extent Panel B, illustrate non-random noise in the estimates. Specifically, the estimates jump up and down in sequence. This curious pattern is due to the fact that ounce multiples alternate on the left and right side of 100-gram multiples and because lower socioeconomic status children are more likely to have birth weights recorded in ounces.[18] In addition, more often than not, the estimates suggest that those just to the left of these cutoffs are of lower socioeconomic status.

As another important test of composition bias, panels C and D examine Apgar scores. For the reasons discussed in the previous section, Apgar scores are not necessarily a good zero test for the 1500-gram threshold since low birth weight classification could conceivably affect Apgar scores.[19] However, there is no reason to expect this to be the case at most other thresholds that are 100-gram multiples. Around these other thresholds, we can say with relative confidence that Apgar scores should vary smoothly along the birth weight distribution in the absence of composition bias. However, these estimates show that children

[18]This pattern is even more evident when a smaller bandwidth is chosen. These results are available upon request.

[19]This is why ADKW do not include this variable in their analysis of covariates.

just below 100-gram thresholds have higher average Apgar scores (Panel C) and a lower chance of being born with an Apgar score below 3 (out of a 10 point scale). It is notable that very few of the estimates are significant at the 95 percent level but the set of estimates as a whole provides conclusive evidence of composition bias.

Figure 8 presents the "donut RD" analog to Figure 7, dropping observations with birth weights recorded in 100-gram and ounce multiples. In large part, the point estimates for Apgar score and other covariates appear more random, highlighting the usefulness of this approach.

5 Additional Examples

In the above example, the limited precision with which birth weights are measured leads to a distribution of the running variable characterized by heaping at 100-gram and ounce multiples. Because low socioeconomic status infants are disproportionately observed at these heaps, underlying characteristics related to mortality outcomes are not a smooth function of the running variable. To the extent to which these heaps fall close to any RD threshold under consideration, estimated effects will be biased.

Although we have focused on birth weights as an example of a potentially problematic running variable throughout much of this paper, there are many other circumstances where compositional changes are likely to occur at data heaps which could hinder valid identification. In this section, we offer two additional examples.

5.1 Day of Birth as a Running Variable

A common approach used to estimate 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, since "five years old on September 1st" is a common school-entry requirement, it is

typical for researchers to compare the outcomes of individuals born just before September 1st who begin school earlier, and thus tend to obtain more years of education, to the outcomes of individuals who are born just after September 1st.

Suppose we were to conduct such a study focusing on those born within a few days before and after September 1st. Even if education and school starting age have no impact on children's outcomes, it is quite likely that we would observe systematically different outcomes for children born just after September 1st relative to those born before.

In the distribution of birthdays around September 1st in 2001, for example, heaping is quite evident as relatively few children are born September 1st through September 3rd. It turns out that these three days coincide with a Saturday and Sunday followed by Labor Day. So why would we expect to see differences in children's outcomes regardless of whether years of education or school-starting age affect outcomes? Because hospitals tend not to schedule induced labor and cesarian sections on weekends (Dickert-Conlin and Elder, forthcoming). As such, children born without medical intervention, who are disproportionately low-socioeconomic status (Dickert-Conlin and Elder, forthcoming), will be over-represented among those just to the right of the cutoff. As a result, any comparison of outcomes is likely to be subject to composition bias.[20]

Again this is a case in which testing the extent to which covariates predict the observed heaping would signal that the identification strategy is inappropriate. Also note that a conventional test for discontinuities in observable characteristics might not identify this problem if the weekend days are pooled with subsequent weekdays to the right of the cutoff, or if many birth cohorts are pooled together. Further, composition bias would remain even if the weekend did not perfectly coincide with the school-entry-date cutoff.[21]

[20]One would expect 29 percent of births to fall on weekends if the distribution were smooth. Yet, only 21 percent of births occurred on weekends in 2001, with 21 percent of weekday deliveries being for nonwhites, and 23 percent of weekend births being for nonwhites. In 2001, 21 percent of mothers giving birth on weekdays had less than a high school education and 23 percent of mothers giving birth on weekends had less than a high school education.

[21]Note that McCrary and Royer (forthcoming) – which is the only paper we are aware of using exact dates

5.2 Self-reported Age as a Running Variable

When pension eligibility is determined by age, in the absence of panel data, a regression discontinuity design provides a natural alternative to estimating its effects on individual outcomes. Essentially, such an approach compares the outcomes of individuals with ages just above the eligibility cutoff to the outcomes of those with ages just below.

In circumstances in which individuals may not know their exact ages, this could be problematic. For example, it is common to observe heaping at decades in self-reported-age data from developing countries. Further, individuals reporting their ages in decades tend to look different from similarly-aged individuals who do not report their ages in decades (Edmonds, Mammen, and Miller 2005). If an age-interval of ten falls closer to one side of eligibility cutoff than the other, it would likely bias the RD estimated effect of pension eligibility.

6 Discussion and Conclusion

The results we have presented suggest that there is still much work to be done in documenting the efficacy of treatment provided to at-risk newborns. In contrast to prior work, when we focus on a sample unlikely to be subject to composition bias, our analysis reveals no significant benefits to very low birth weight classification. If very low birth weight classification is truly tied to more intensive treatment then it appears to not translate into significant declines in infant mortality.

More generally, in this paper we have raised an important concern for researchers that employ RD designs. Specifically, estimated effects are likely to be subject to composition bias when attributes related to the outcomes of interest predict heaping in the running variable.

of birth in this context – adopt a wide bandwidth and pool across a large number of cohorts, both actions acting to mitigate the bias that would otherwise result.

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 treatment. In this paper, we highlight the fact that abrupt compositional changes are not uncommon in a wide variety of data as a result of non-random heaping. In these circumstances, the standard RD assumption that characteristics related to outcomes vary smoothly with the running variable is unlikely to hold and thus estimated effects are likely to be biased. Because heaps are often pooled with adjacent data points, conventional tests may fail to identify important compositional differences on either side of a cutoff. Further, conventional tests are especially likely to fail to identify compositional differences when heaping does not perfectly coincide with the treatment threshold under consideration.

For this reason, we propose a more rigorous approach to establishing the validity of RD designs when the distribution of the running variable has reporting heaps. Regardless of whether or not the heaps fall directly on one side of the treatment threshold, composition changes at these heaps can bias estimated treatment effects. As such, it is important to test whether the heaps are predicted by underlying characteristics, as in Figure 3.

In addition to heaping in the distribution of the running variable, our results point to two other phenomena that can signal that this type of composition bias may be of concern. First, non-random heaping leads to estimated effects where they are not expected (Figure 5). Second, these estimated effects are quite sensitive to the choice of bandwidth (compare Figure 5 and Figure A5). Third, these estimated effects are sensitive to the addition of covariates (compare panels C and D of Table 1).

A straightforward way to deal with this problem, after being diagnosed, is to restrict the sample in a manner that balances covariates across the threshold. In the birth weight example

above, we accomplish this goal by dropping observations coinciding with heaps in the running variable (i.e., 100-gram and ounce multiples). While this "donut RD" effectively deals with the composition bias resulting from non-random heaping, it can serve as a robustness check for any RD design in which non-random sorting or heaping is a potential concern.

References

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.

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

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

Edmonds, E. V., K. Mammen, and D. R. Miller (2005): "Rearranging the family? Income Support and Elderly Living Arrangements in a Low-Income Country," *Journal of Human Resources*, 40(1), 186–207.

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

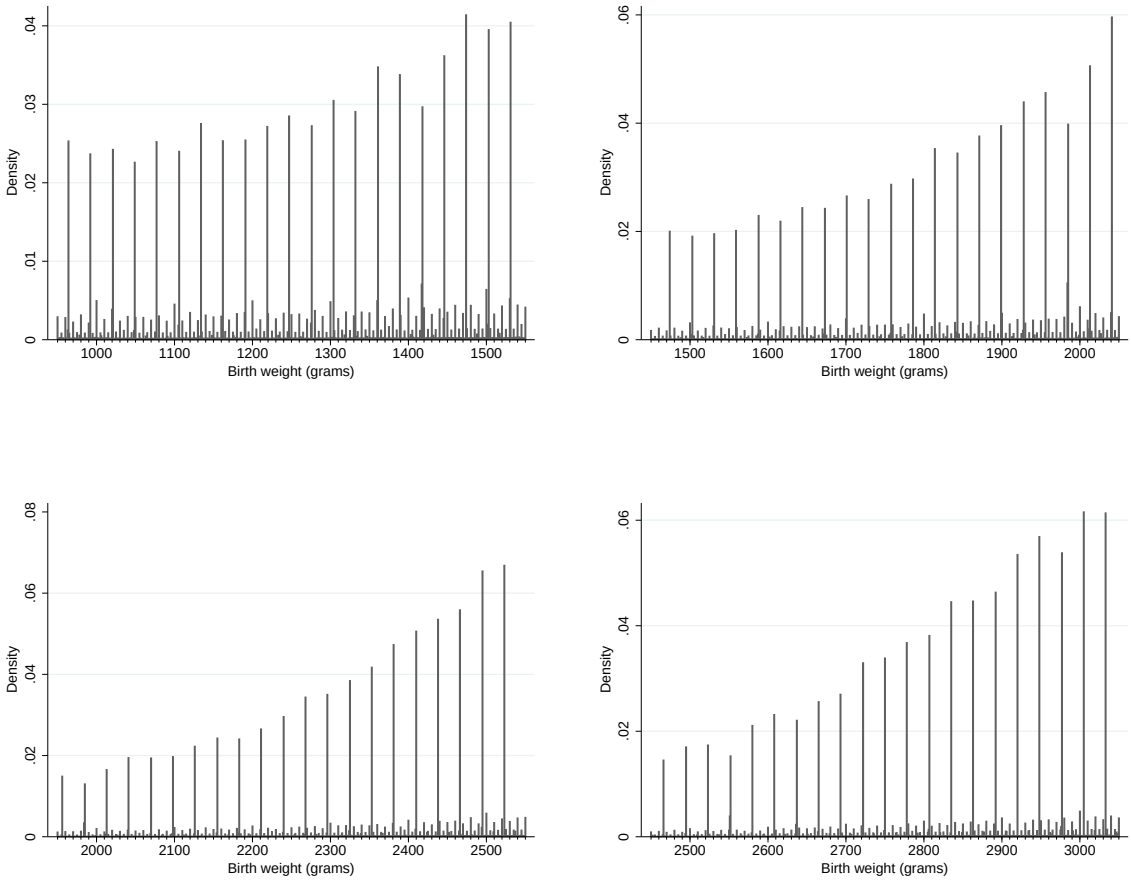
Lee, D. S., and D. Card (2008): "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 127(2), 655–674.

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 (forthcoming): "The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth," *American Economic Review*.

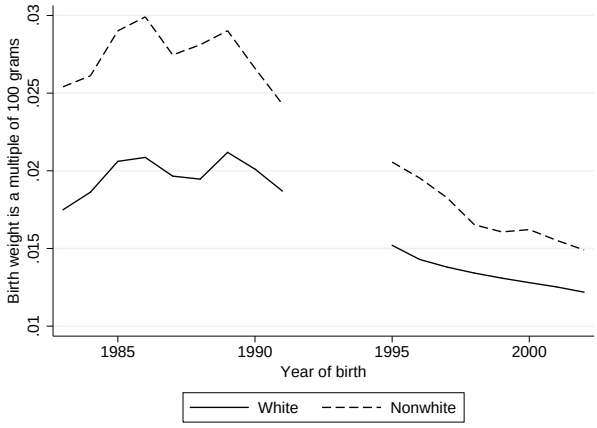
Figure 1
Distribution of Birth Weights



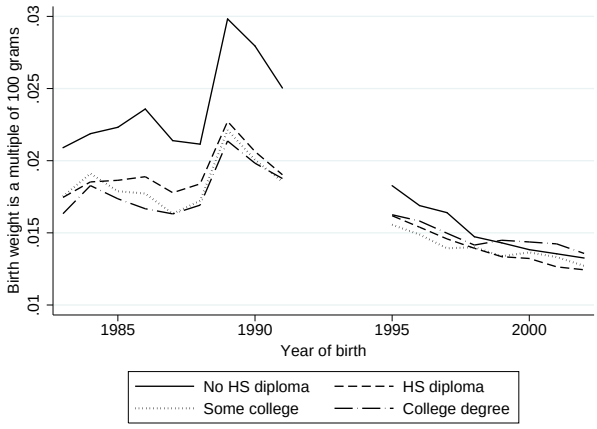
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure 2
Fraction of Births Recorded in 100s of Grams Over Time

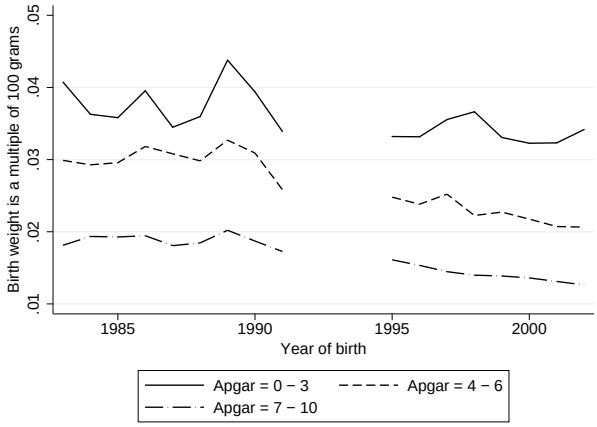
Panel A: By Race



Panel B: By Mother's Education

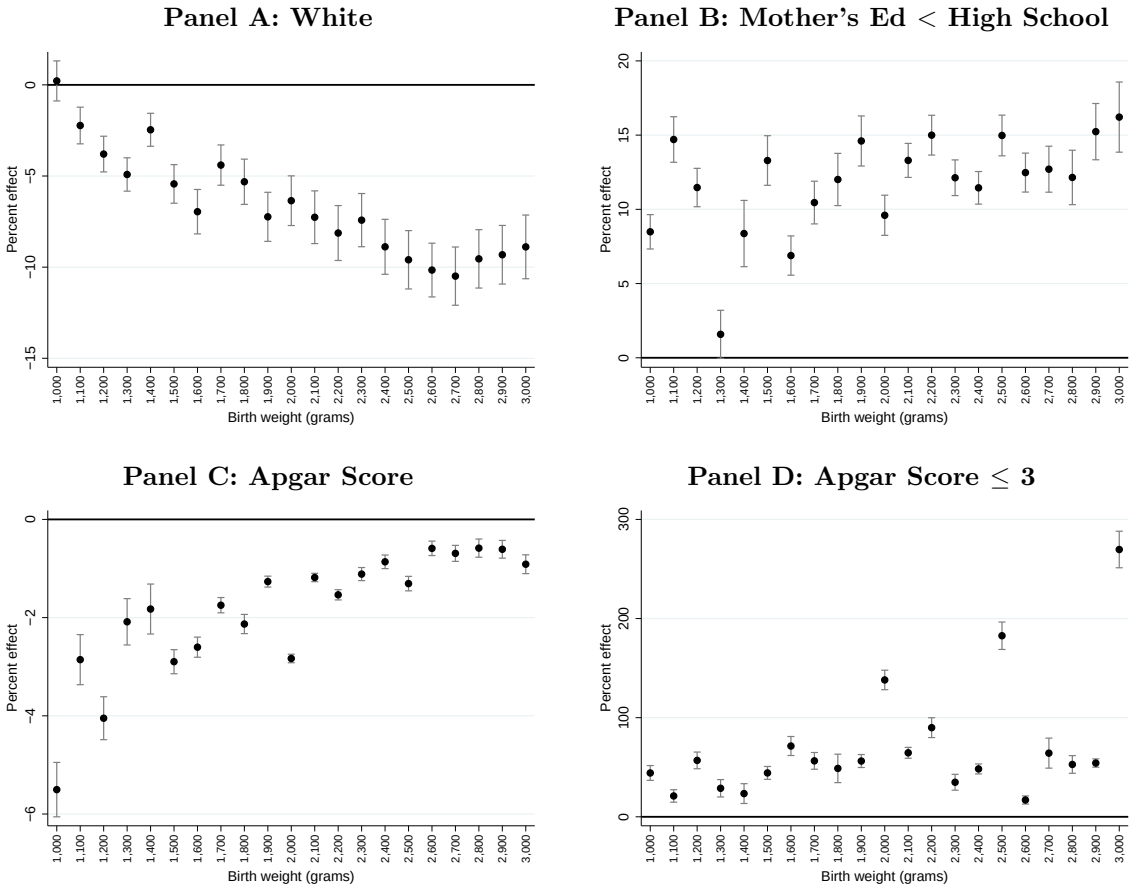


Panel C: By Apgar Score



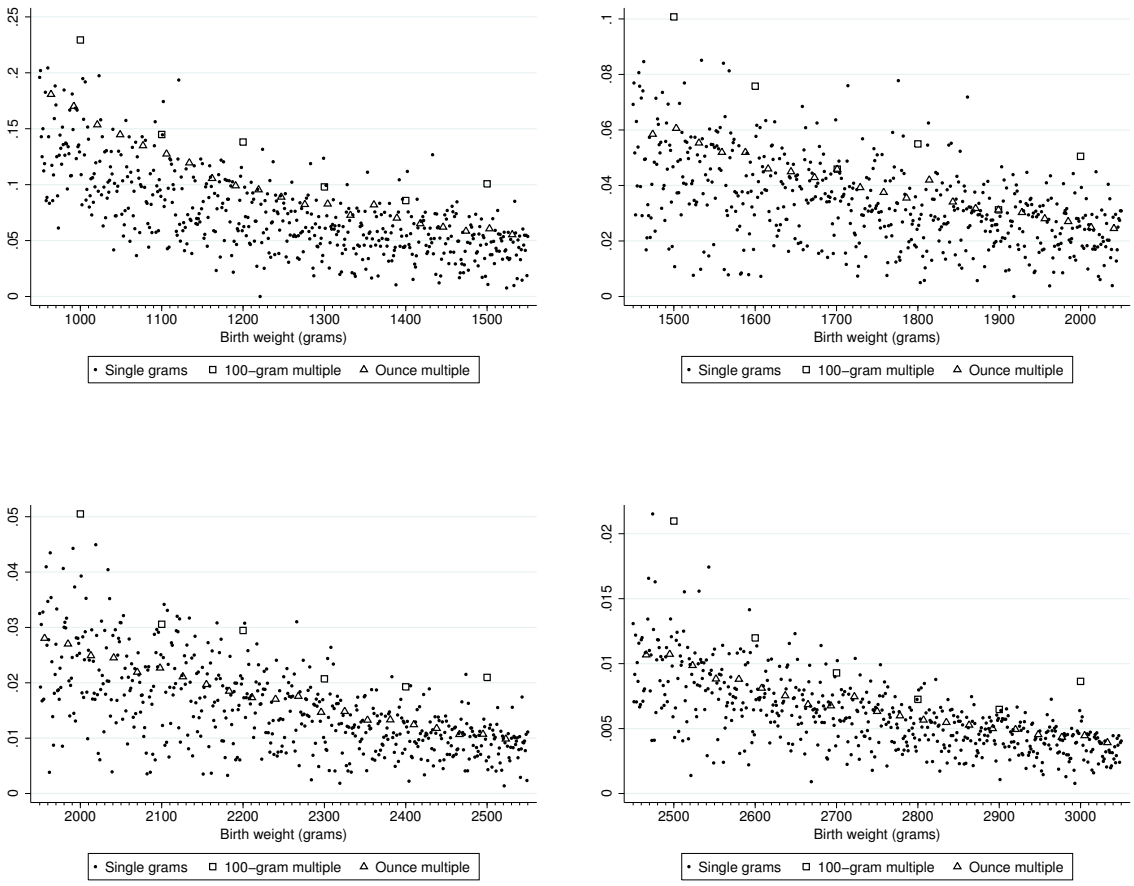
Note: Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure 3
Estimated Jumps in Child Characteristics at 100-Gram Multiples



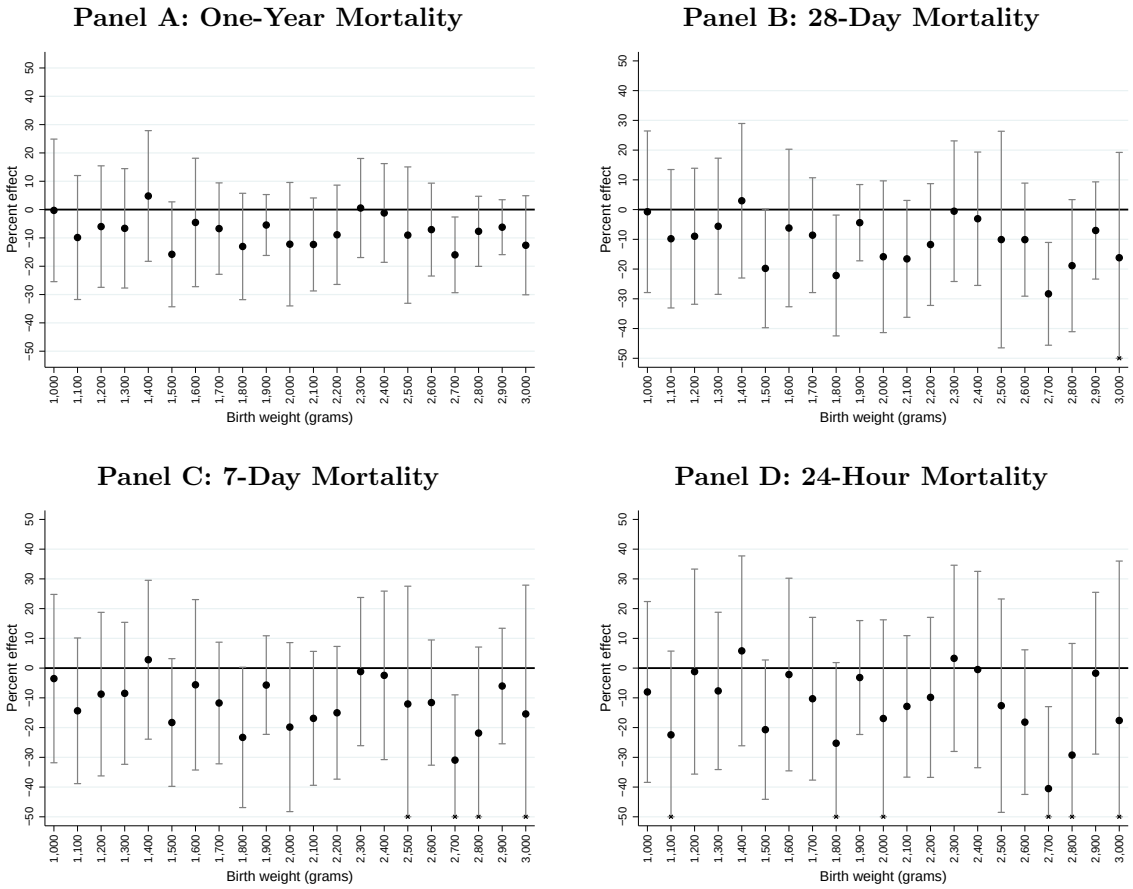
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure 4
Mean of One-Year Mortality Probabilities



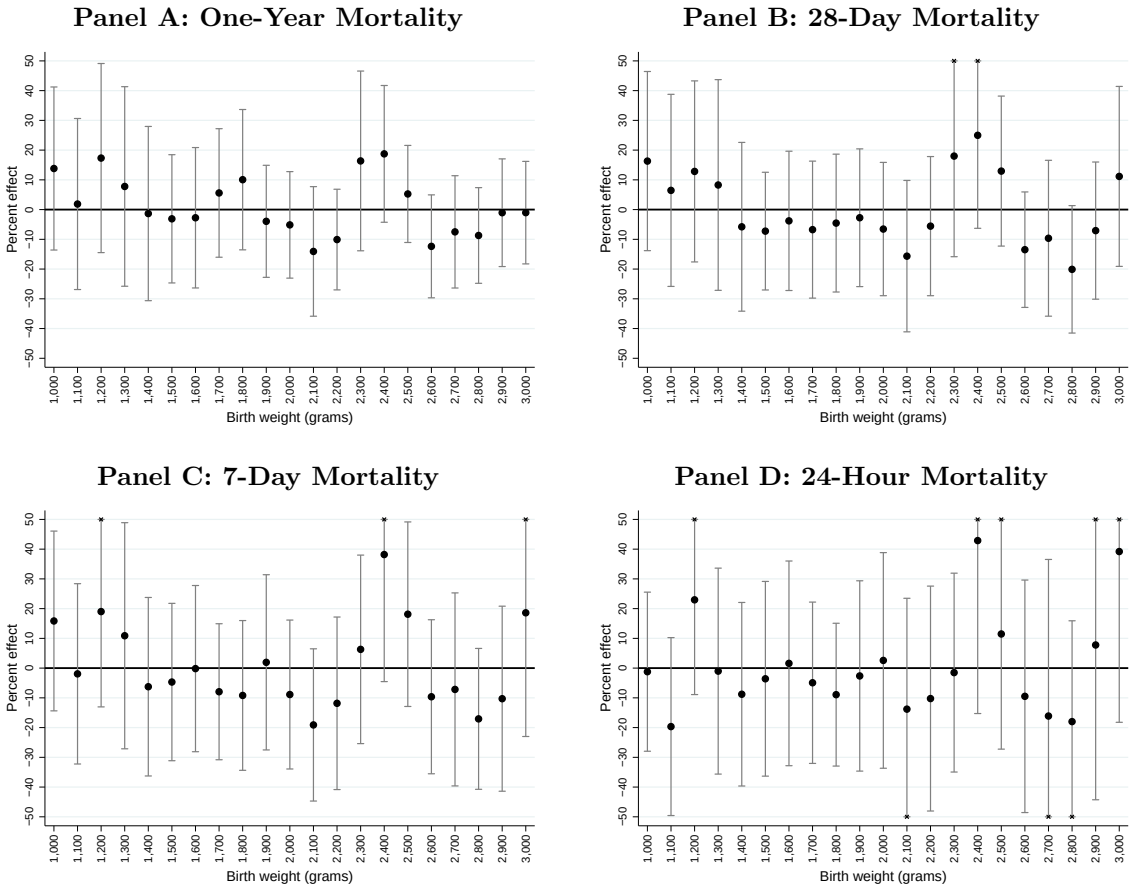
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure 5
Estimated Impacts of Having Birth Weight $< Z$



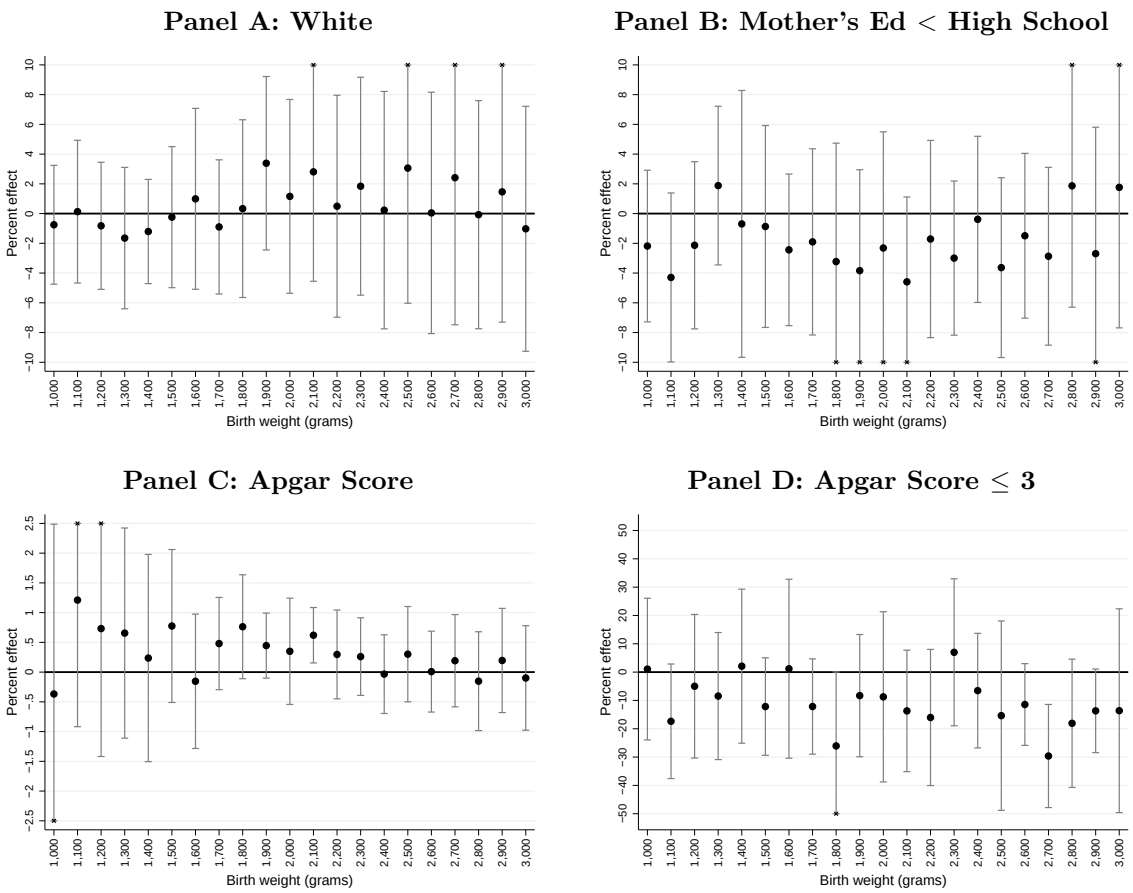
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure 6
Donut RD Estimated Impacts of Having Birth Weight $< Z$



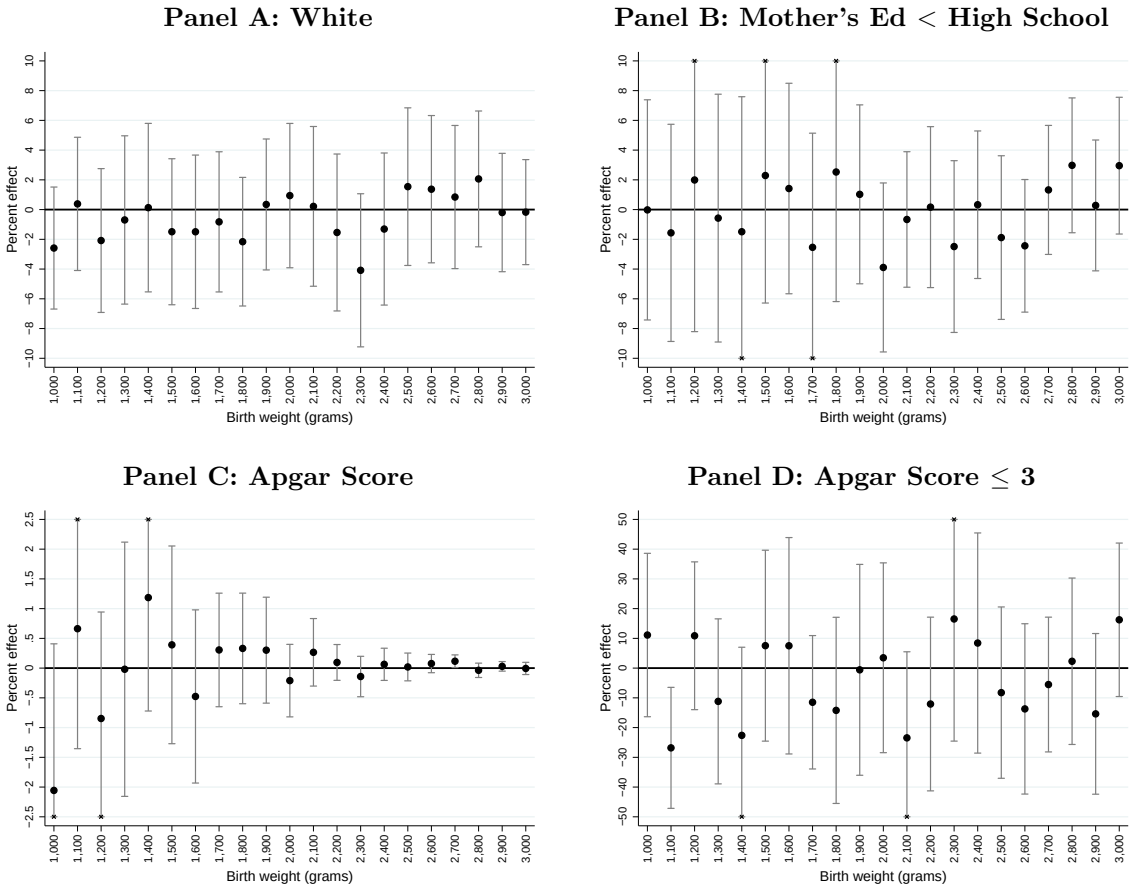
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). Children with birth weights recorded in 100s of grams or in ounces are not included in the analysis sample.

Figure 7
Standard Tests for Discontinuities in Characteristics



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

Figure 8
Donut RD Tests for Discontinuities in Characteristics



Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). Children with birth weights recorded in 100s of grams or in ounces are not included in the analysis sample.

Table 1
Replication of ADKW's Main Results Along With Donut RD Estimates

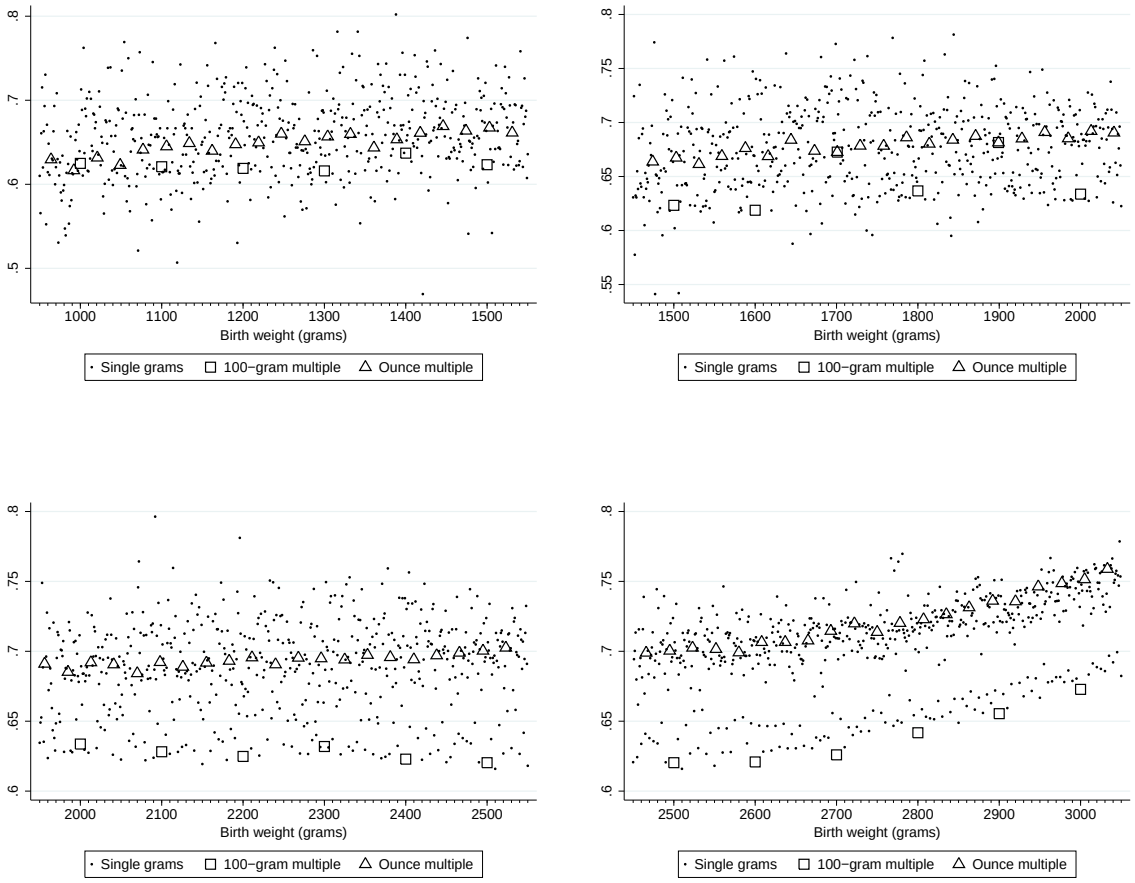
Mortality Outcome	One-Year (1)	28-Day (2)	7-Day (3)	24-Hour (4)
Panel A: ADKW's estimates				
Weight < 1500 grams	-0.0095*	-0.0088*	-0.0060	-0.0043
	(0.0048)	(0.0038)	(0.0032)	(0.0023)
Observations	202,071	202,071	202,071	202,071
Panel B: Our replication of ADKW				
Weight < 1500 grams	-0.0095*	-0.0088*	-0.0060	-0.0042
	(0.0047)	(0.0037)	(0.0032)	(0.0023)
Observations	202,078	202,078	202,078	202,078
Panel C: Estimates using ADKW's control variables				
Weight < 1500 grams	-0.0071	-0.0071*	-0.0046	-0.0033
	(0.0041)	(0.0032)	(0.0028)	(0.0020)
Observations	202,078	202,078	202,078	202,078
Panel D: Estimates using ADKW's control variables and Apgar scores				
Weight < 1500 grams	-0.0038	-0.0040	-0.0018	-0.0004
	(0.0025)	(0.0021)	(0.0018)	(0.0011)
Observations	159,315	159,315	159,315	159,315
Panel E: Estimates controlling for 100-grams				
Weight < 1500 grams	-0.0054	-0.0057*	-0.0035	-0.0026
	(0.0031)	(0.0026)	(0.0022)	(0.0016)
Observations	202,078	202,078	202,078	202,078
Panel F: Donut RD dropping those at 100-grams				
Weight < 1500 grams	-0.0054	-0.0057*	-0.0035	-0.0026
	(0.0031)	(0.0026)	(0.0022)	(0.0016)
Observations	198,534	198,534	198,534	198,534
Panel G: Estimates controlling for 100-gram and ounce multiples				
Weight < 1500 grams	0.0003	-0.0005	-0.0000	-0.0001
	(0.0050)	(0.0033)	(0.0032)	(0.0026)
Observations	202,078	202,078	202,078	202,078
Panel H: Donut RD dropping those at 100-gram and ounce multiples				
Weight < 1500 grams	0.0000	-0.0007	-0.0001	-0.0002
	(0.0050)	(0.0033)	(0.0032)	(0.0026)
Observations	53,974	53,974	53,974	53,974

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. No controls are included except where noted. The controls referred to in Panel C and Panel D include measures of prenatal care, mother's age, mother's education, father's age, child gender, gestational age, mother's race, plurality of birth, birth order, and year. In Panel D, Apgar scores enter the model as fixed effects. In panels E and G, observations at reporting heaps are controlled for with unrestricted fixed effects.

* significant at 5%; ** significant at 1%.

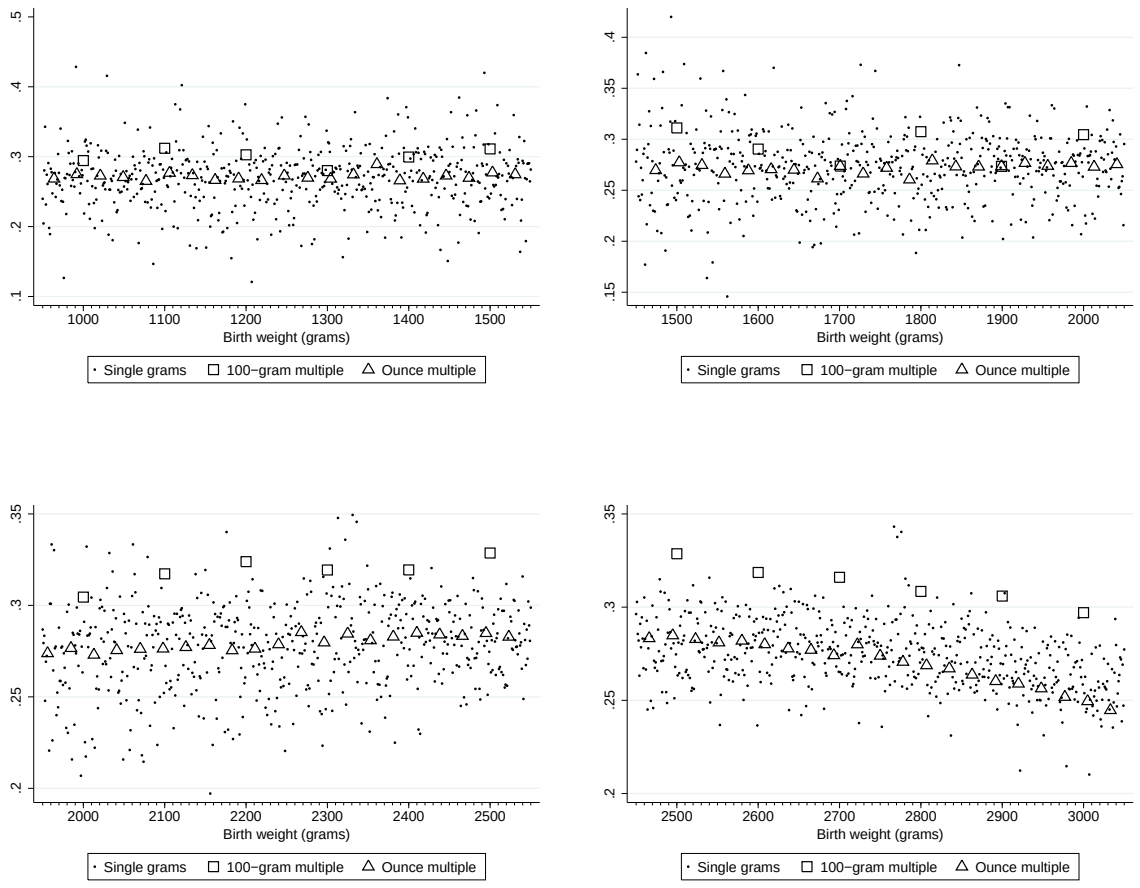
Appendix: Additional Figures

Figure A1
Fraction White



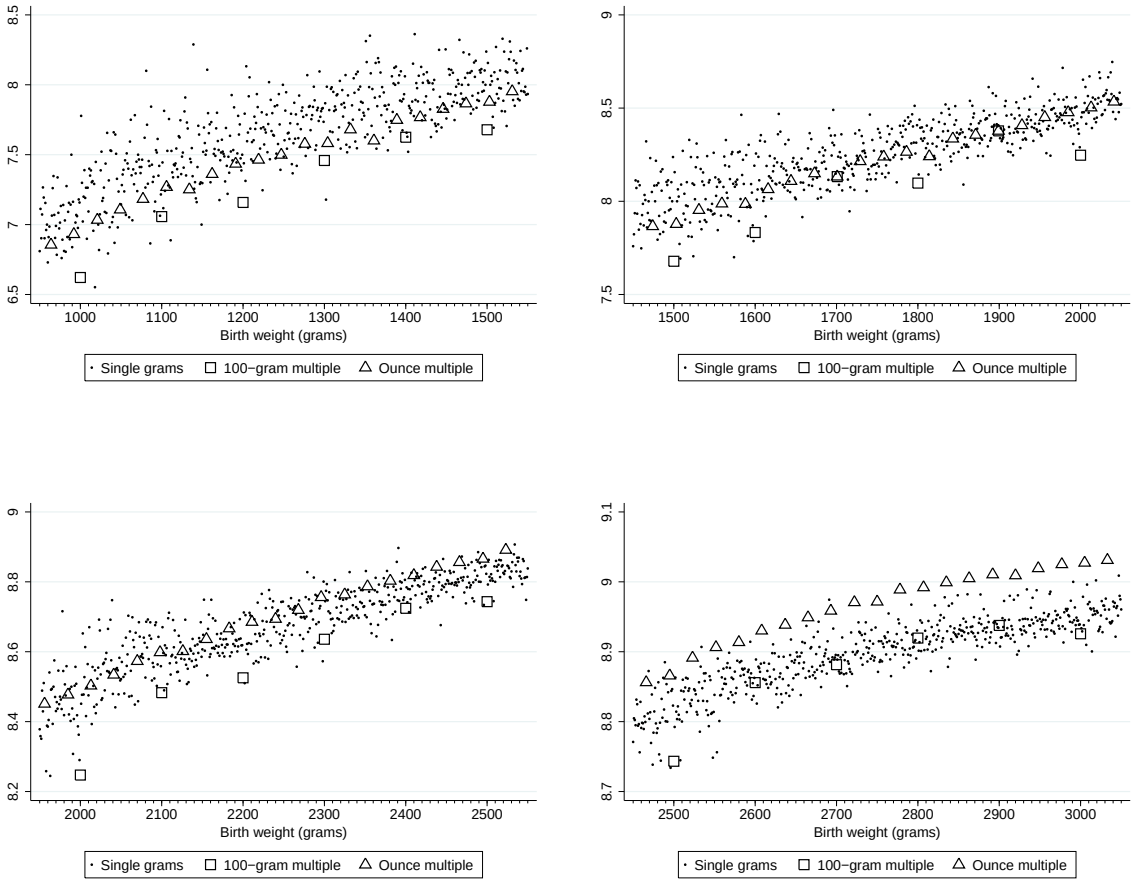
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure A2
Fraction of Mother's Education With Less than a High School Education



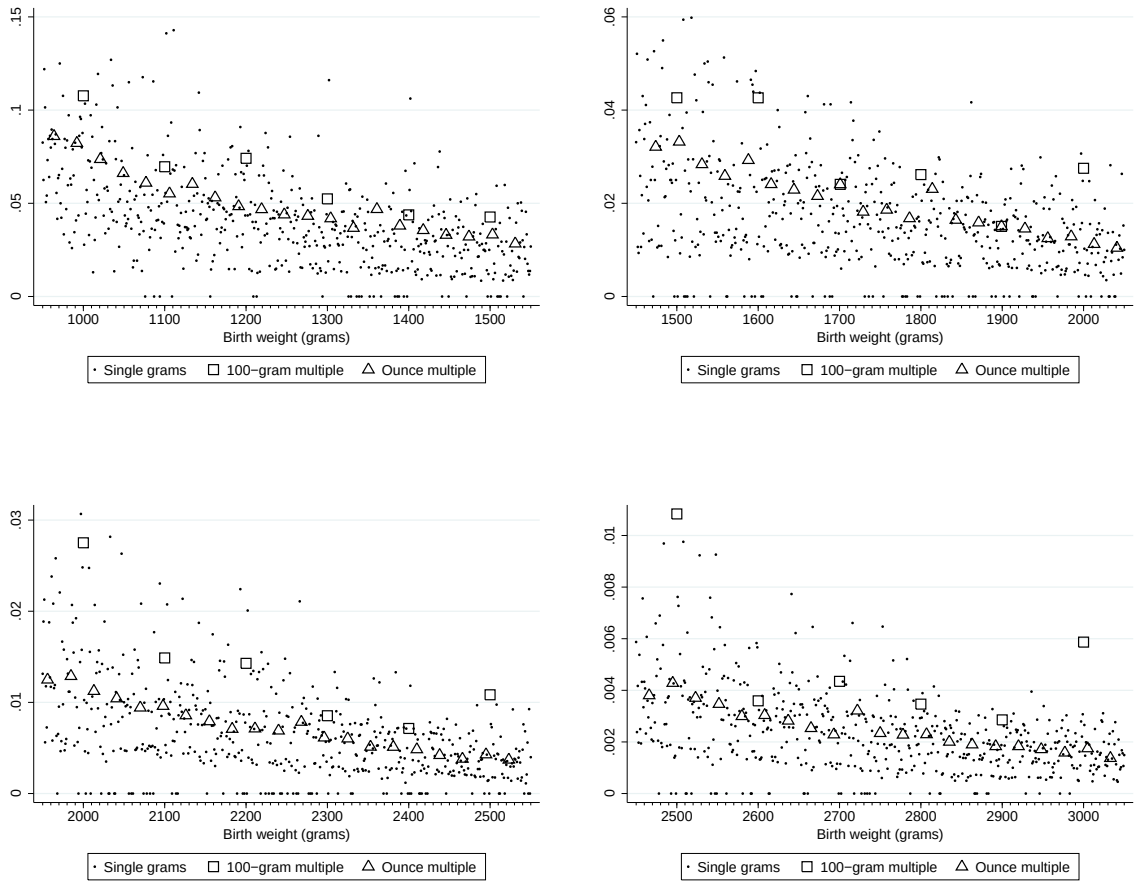
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure A3
Mean of Apgar Scores



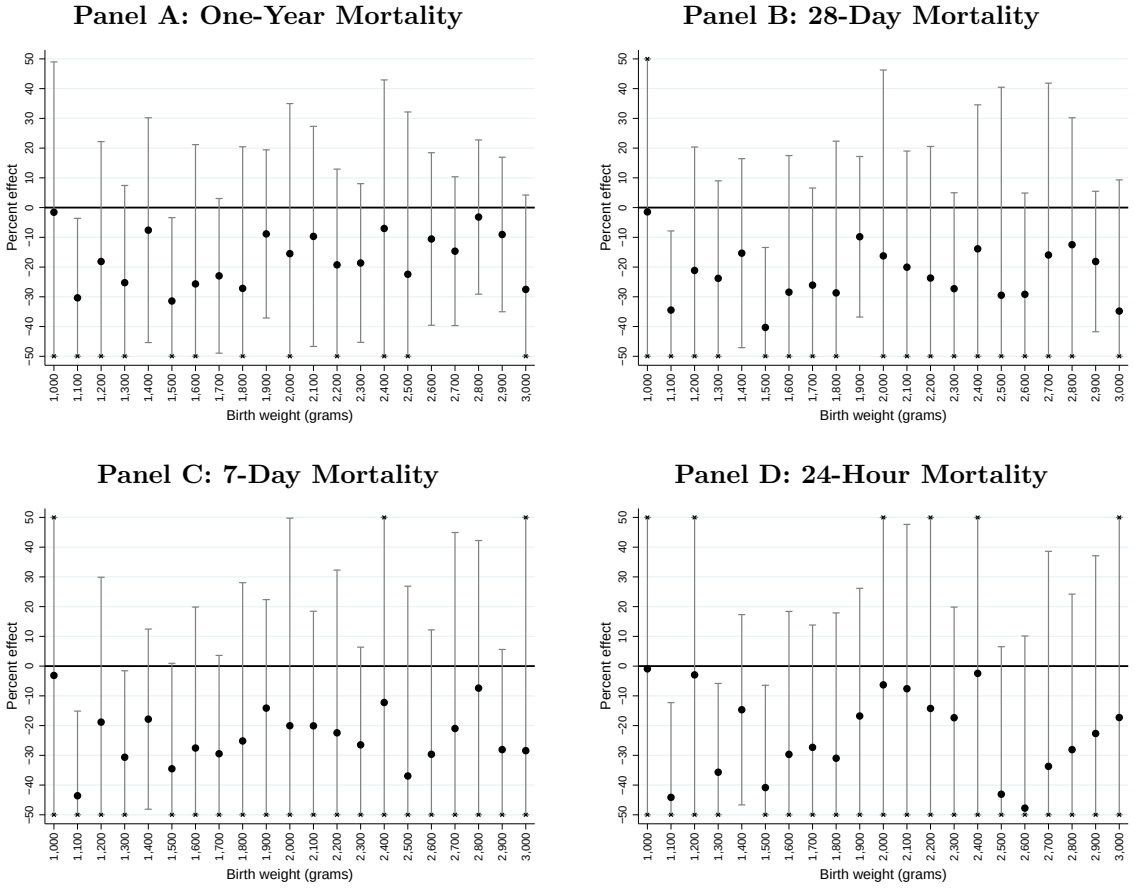
Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994).

Figure A4
Fraction With Apgar Score < 3



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

Figure A5
Bandwidth Sensitivity for Estimated Impacts of Having Birth Weight $< Z$



Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). Estimates using a bandwidth of 30 grams are shown with their 95% confidence intervals.