SAVING BABIES? REVISITING THE EFFECT OF VERY LOW BIRTH WEIGHT CLASSIFICATION*

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

The Quarterly Journal of Economics, *forthcoming*

Abstract

We reconsider the effect of very low birth weight classification on infant mortality. We demonstrate that the estimates are highly sensitive to the exclusion of observations in the immediate vicinity of the 1,500-gram threshold, weakening the confidence in the results originally reported in Almond, Doyle, Kowalski, and Williams (2010).

*The authors thank Robert Barro, Todd Elder, Hilary Hoynes, Thomas Lemieux, Justin McCrary, Doug Miller, Marianne Page, Heather Royer, Larry Singell, and Ann Huff Stevens for their comments and suggestions.

In Barreca, Guldi, Lindo, and Waddell (2010) we highlight various econometric issues related to regression discontinuity designs in which there is heaping in the running variable. Motivated by this analysis, we reconsider a recently published result with far-reaching policy implications that is not robust to the issues raised therein. In particular, Almond, Doyle, Kowalski, and Williams (2010), henceforth ADKW, use a regression discontinuity design to argue that one-year infant mortality decreases by approximately one percentage point as birth weight crosses the 1,500-gram "very-low-birth-weight (VLBW)" threshold from above. Relative to mean one-year mortality of 5.5 percent just above 1,500 grams, the implied treatment effect is sizable, suggesting large returns to promoting the types of medical interventions triggered by VLBW classification. Given the importance of the research question, we reconsider the point estimate derived around this VLBW threshold.

ADKW's analysis follows standard regression discontinuity practices. They show the sensitivity of the results to different bandwidth choices, to the inclusion of a large number of control variables, and test whether observable characteristics are discontinuous through the VLBW threshold. They also consider the distribution around the threshold, as excess mass on one side or the other would raise concern that individuals might manipulate their recorded weights in order to receive favorable treatment. Their investigation revealed extensive heaping at one-ounce and 100-gram multiples, which can also be explained by technological constraints in measurement precision and natural tendencies to round to round numbers for convenience. In an effort to argue that the heaping around the 1,500-gram threshold is "not irregular" and hence not of concern, they argue that similar heaps are found around 1,400 grams and 1,600 grams where individuals would have no incentive to act in a strategic manner. Using McCrary's (2008) estimation strategy, they also appeal to the lack of a statistically significant estimate of the discontinuity in the distribution.

Nevertheless, it turns out that the 1,500-gram heap *is* irregular in a critical fashion. In particular, those at this data heap have substantially higher mortality rates than surrounding observations on *either* side of the VLBW threshold. This feature of the data is demonstrated in Figure I, in which we illustrate unadjusted mean mortality rates across the distribution of birth weights around 1,500 grams.[1] In each of the four panels, documenting 24-hour through one-year mortality rates, those at the 1,500-gram heap appear to be of a particularly disadvantaged sort. They are an outlier with respect to both those on the left of 1,500 grams *and* those on the right of 1,500 grams. This may be a signal that poor quality hospitals have relatively high propensities to round birth weights but is also consistent with manipulation of recorded birth weights by doctors, nurses, or parents in order to obtain favorable treatment for their children.[2] In Barreca, Guldi, Lindo, and Waddell (2010), we show that this non-random heaping leads one to conclude that it is "good" to be strictly less than *any* 100-gram cutoff between 1000 and 3000 grams.[3]

Regardless of the mechanism, it raises the concern that ADKW's estimates may be driven in large part by the outliers at the 1,500-gram heap.[4] Given that the motivation for regression discontinuity designs is a comparison of means as estimates *approach* the treatment threshold from each side, estimates should not be sensitive to the observations at the threshold.[5]

Motivated by this concern, we consider a "donut RD" of sorts, whereby we systematically

[1]Note that our Figure I is a disaggregated version of Figure II in ADKW.

[2]The direction of the abrupt change at the 1,500-gram heap is consistent with healthier types non-randomly sorting to the left of the cutoff. It is important to note more generally that there tend to be systematic compositional changes at *all* 100-gram and ounce multiples (Barreca, Guldi, Lindo, and Waddell 2010). However, the concern about manipulation rather than something more benign driving the abrupt change is more worrisome for the 1,500-gram heap as it is even more of an outlier than the 1,400 and 1,600-gram heaps, as shown in Figure I. In their reply, ADKW seem to have misunderstood that there are reasons to be concerned with manipulated observations ending up on the left of the cutoff despite the heaping observed to the right of the cutoff. Moreover, given a potential incentive to manipulate birth weights downward, the essential question is which attributes predict manipulation. As they mention in their reply, "newborns at exactly 1,500 grams are anomalous based on *ex ante* characteristics such as race and mother's education." To be more specific, Barreca, Guldi, Lindo, and Waddell (2010) demonstrate that they are substantially less likely to be white and more likely to have a mother with less than a high school education.

[3]ADKW mention having done this same robustness check, reporting that the "results support the validity of [their] main findings." We disagree with this interpretation of the results. Using a bandwidth of 85 grams 37 of 41 placebo estimates indicate that mortality is lower to the left of the cutoff. With a bandwidth of 30 grams, 41 of 41 placebo estimates indicate that mortality is better to the left of the cutoff.

[4]Although ADKW report that "the results are qualitatively similar across a wide range of bandwidths," note that their mortality estimates more than triple across the bandwidths used in their sensitivity analysis, which is consistent with this concern.

[5]This is the statistical argument, expanded upon in the next paragraph, that ADKW choose to discount completely when they write "there is no general economic or statistical case for exclusion of observations at or around the threshold in a regression discontinuity (RD) design" in their reply.

remove observations in the immediate vicinity of the 1,500-gram heap and re-estimate the discontinuity on the remaining sample. In doing so, we continue to compare mean estimates as they approach the VLBW threshold from each side, while allowing for the possibility that those at the heap are systematically different from surrounding observations. By expanding the size of the "donut hole" to include more than just the 1,500-gram threshold itself, the approach further addresses potential concerns that there is non-random sorting around the VLBW threshold. That said, as we consider dropping those falling exactly at the cutoff, and then those within one, two, or three grams of the cutoff, it is worth recognizing just how incremental these considerations are. For example, even under our most extreme sample restriction the implied gap in birth weights between the observations to the left and right of the cutoff is roughly equivalent in weight to seven paper clips (i.e., seven grams). Given that the baseline birth weight in consideration is 1,500 grams, or roughly the weight of Simon and Blume's textbook *Mathematics for Economists*, this seems a reasonable accommodation given the concerns described above. Again, if the underlying identification strategy is valid, we anticipate that estimates will be robust to such a restriction. If the results are shown to be sensitive, however, it calls the identification into doubt.[6]

With this in mind, in Panel A of Table 1 we report the estimates of our replication of ADKW.[7] We then begin our sensitivity analysis by estimating the effects after dropping those falling exactly at the 1,500-gram heap, shown in Panel B. This very small sample restriction, which only reduces the sample size by approximately two percent and removes only one cluster, causes the estimated impact on one-year mortality to fall by more than 50 percent.[8]

[6] We find it odd that ADKW, in their reply, support the use of the donut RD as "a useful robustness check that [they] should have included in [their] original paper" yet are resistant to increasing the size of the hole, stating that they "see no clear case for excluding the larger set of newborns from 1,497 to 1,503 grams." The statistical argument supporting a donut RD with a hole of any size is the same, whether extremely small (1 gram or 0.07 percent of the cutoff weight) or less-extremely small (7 grams or 0.47 percent of the cutoff weight).

[7] These estimates are identical to those presented by ADKW although we have seven additional observations.

[8] In the subsequent analysis discussed below, dropping observations within one gram of the VLBW threshold removes an additional 0.001 percent of observations, dropping observations within two grams of the

In panels C through E we drop observations within one, two and three grams of the VLBW threshold, respectively. This series of estimates casts further doubt on the previously published conclusions, as they become smaller when those within one gram of the cutoff are omitted, and smaller still when those within two grams of the cutoff are omitted. Finally, we omit those within three grams of the cutoff, which reduces the sample by 11 percent because 1,503 grams corresponds to a large data heap at 53 ounces—an additional source of potential bias.[9] With this restriction, the point estimate is now 25 percent of the published estimate and statistically indistinguishable from zero. Overall, this collection of estimates substantially weakens the confidence in the results originally reported by ADKW, and highlights the importance of considering the fuller implications of heaping in running variables, as explored more generally in Barreca, Guldi, Lindo, and Waddell (2010).

In this issue, ADKW conduct a donut RD analysis like that described above. Their main results now focus on children born in "low-level neonatal intensive care hospitals" in California. This sample consists of only twenty-two percent of the sample of California births (omitting those born in "high-level neonatal intensive care hospitals") where data is linked to hospital quality.[10] It consists of only thirteen percent of all children they can link to hospital costs (omitting children born in Arizona, Maryland, New York, and New Jersey despite the fact that their original work shows a larger first-stage effect of VLBW classification on medical care for the "five-state sample" than for California). It includes less than two percent of the children whose birth weights are linked to mortality outcomes. After making all of these data restrictions, it appears as if they have found a setting that provides some

threshold removes an additional 0.006 percent of observations, and dropping observations within three grams of the threshold removes an additional 11 percent of observations. The final restriction reduces the sample size by a larger degree because 1,503 grams corresponds to 53 ounces, where there is a large data heap. It is worth noting, however, that each of these restrictions only removes two additional clusters of data. For an in-depth discussion of the importance of recognizing correlation within clusters in RD designs when the running variable is discrete, see Lee and Card (2008).

[9] In this issue, ADKW concede that the exclusion of those at the 1,500-gram heap is a useful robustness check. Yet, they deny the logical imperative that the exclusion of those at the 1,503-gram heap should likewise be useful.

[10] We also note that ADKW's analysis of hospital costs in California presented in this issue use 16,528 observations whereas they used only 14,560 observations in their original work (Table A6).

evidence to support their hypothesis. However, even after choosing this extremely narrow subsample, their first stage is fragile which casts further doubt on these results as being informative about the marginal returns to hospital care. The final set of donut RD estimates they present does not indicate a significant effect of very low birth weight classification on hospital costs.[11] Further, we note that their first stage does not lose significance because the donut RD sample restrictions increase the standard error estimate but because their coefficient estimate falls by fifty-eight percent. As such, we disagree with their assertion that their results are "robust."

TULANE UNIVERSITY

MOUNT HOLYOKE COLLEGE

UNIVERSITY OF OREGON and IZA BONN

UNIVERSITY OF OREGON and IZA BONN

References

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

— , — , — , and — , "The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: A Reply to Barreca, Guldi, Lindo, and Waddell," (This Issue).

Barreca, Alan, Melanie Guldi, Jason M. Lindo, and Glen R. Waddell, "Running and Jumping Variables in Regression Discontinuity Designs," *Mimeo*, (November 2010).

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

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

Simon, Carl P. and Lawrence Blume, *Mathematics for Economists* (W. W. Norton & Company, Inc., 1994).

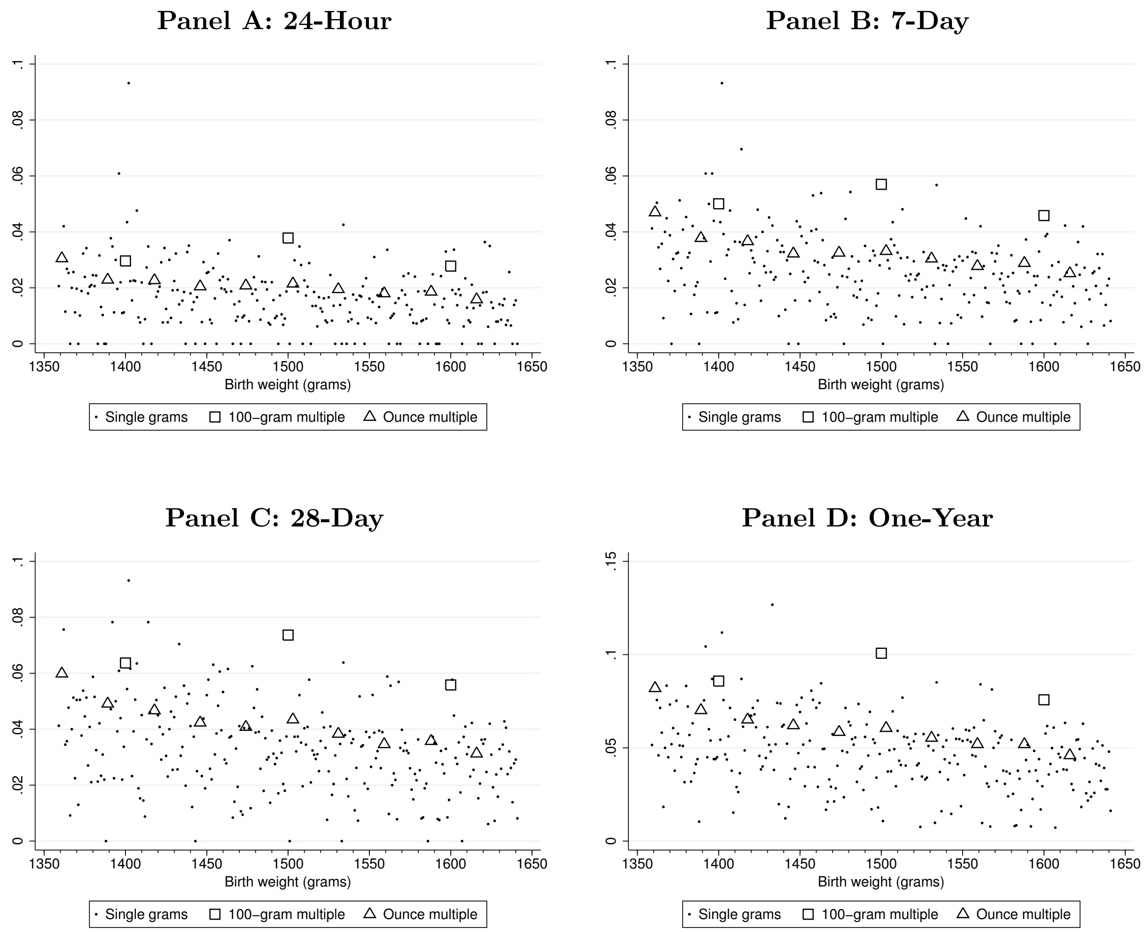
[11] They do not provide a partial F-statistic but the p-value on the estimated discontinuity in hospital costs is 0.37.

Table I
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: Our replication of ADKW's estimates				
Weight < 1,500 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
Clusters	171	171	171	171
Panel B: Donut RD dropping those at 1,500-grams				
Weight < 1,500 grams	-0.0033*	-0.0042**	-0.0023	-0.0018
	(0.0014)	(0.0013)	(0.0013)	(0.0010)
Observations	198,534	198,534	198,534	198,534
Clusters	170	170	170	170
Panel C: Donut RD dropping those within 1 gram of 1,500-gram cutoff				
Weight < 1,500 grams	-0.0035*	-0.0043**	-0.0024	-0.00183
	(0.0014)	(0.0012)	(0.0013)	(0.0010)
Observations	198,334	198,334	198,334	198,334
Clusters	168	168	168	168
Panel D: Donut RD dropping those within 2 grams of 1,500-gram cutoff				
Weight < 1,500 grams	-0.0027*	-0.0037**	-0.0019	-0.0013
	(0.0014)	(0.0012)	(0.0012)	(0.0009)
Observations	197,135	197,135	197,135	197,135
Clusters	166	166	166	166
Panel E: Donut RD dropping those within 3 grams of 1,500-gram cutoff				
Weight < 1,500 grams	-0.0018	-0.0026	-0.0018	-0.0011
	(0.0019)	(0.0015)	(0.0015)	(0.0014)
Observations	175,108	175,108	175,108	175,108
Clusters	164	164	164	164

Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). 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. All estimates include controls for 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. * significant at 5%; ** significant at 1%.

Figure I
Means of Mortality Rates



Note: Results are based on Vital Statistics Linked Birth and Infant Death Data, United States, 1983–2002 (not including 1992–1994). The lower panels of this figure are disaggregated versions of ADKW's Figure II.