The Effects of Poor Neonatal Health on Children's Cognitive Development

David Figlio, Northwestern University and NBER
Jonathan Guryan, Northwestern University and NBER
Krzysztof Karbownik, Uppsala University and Uppsala Center for Labor Studies
Jeffrey Roth, University of Florida

February 18, 2013

We are grateful to the Florida Department of Health and Florida Department of Education for providing us access to merged education and health data for the purposes of this project, and for the technical support in interpreting the key variables described herein. Figlio and Roth appreciate the financial support of the National Science Foundation, as well as the U.S. Department of Education and the Bill and Melinda Gates Foundation through the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Karbownik appreciates the support of a Hedelius Scholarship that permitted him to conduct this research while visiting Northwestern University. Seminar participants at CESifo (Munich), Dartmouth College, Emory University, Northwestern University, Roosevelt University, Tulane University, University of California-Irvine, University of Florida, University of Maryland, University of Texas, University of Wisconsin, and Uppsala University provided helpful comments and feedback. All errors of omission and commission are our own.

I. Introduction

A large literature documents the effects of neonatal health (commonly proxied by birth weight) on a wide range of adult outcomes such as wages, disability, adult chronic conditions, and human capital accumulation. A series of studies, conducted in a variety of countries, including Canada, Norway, and the United States, have made use of twin comparisons to show that the heavier twin in the pair is more likely to have better adult outcomes measured in various ways.[1]

While the existing literature makes clear that there appears to be a permanent effect of poor neonatal health on socio-economic and health outcomes, it provides no guidance regarding the potential pathways through which these effects come into being. For instance, we know very little to date about whether the effects of poor neonatal health on cognitive development varies at different ages (say, at kindergarten entry versus third grade versus eight grade), and no existing study identifies whether school quality could help to mitigate the effects of poor neonatal health on cognitive development. For that matter, we know virtually nothing about whether these effects vary heterogeneously across different demographic or socio-economic groups, so it is impossible given the extant literature to know whether early neonatal health and parental inputs are complements or substitutes. As such, while we have strong evidence from twin comparison studies that poor initial health conveys a disadvantage in adulthood, we have virtually no information about the potential roles for policy interventions in ameliorating this disadvantage during childhood.

The reason for these gaps in the literature involves data availability. The datasets that previous researchers have used to study the effects of poor neonatal health on adult outcomes (e.g., Scandinavian registry data, or data matching a mother's birth certificate to her children's birth certificates) do not include information on schooling and human capital measures during key developmental years. And even the small number of studies that investigate the effects of birth conditions on test scores rather than adult outcomes (Bharadwaj et al., 2010; Rosenzweig and Zhang, 2009) are in developing contexts (e.g., Chile and China) that lack the diversity necessary to explore heterogeneous effects of poor neonatal health on cognitive development, or these effects in a western developed context. And while the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) of children born in the United States in 2001 oversamples twins, this data set is too recent to investigate outcomes in late elementary school or adolescence, too small to study heterogeneous effects of birth weight, and does not include cognitive outcomes that have high stakes for children.

Another gap in the adult-outcomes literature is that the subjects of that literature are rather old at present; they were necessarily born in the 1970s and earlier. Given the advances in modern neonatology, it is reasonable to believe that poor neonatal health in the 21st century may bear little resemblance to poor neonatal health fifty years ago.[2] There have been no studies

[1] Examples of influential previous research include Behrman and Rosenzweig (2004) on schooling and wages, Almond et al. (2005) on neonatal outcomes and hospital costs, and Royer (2009) on next generation birth weight, neonatal outcomes and educational attainment, for the United States; Black et al. (2007) on neonatal outcomes, height, IQ, high school completion, employment, earnings and next generation birth weight, for Norway; Oreopoulos et al. (2008) on neonatal outcomes, health outcomes in adolescence, educational attainment and social assistance take up, for Canada; Rosenzweig and Zhang (2012) on educational attainment, wages and weight for height, for China; Torche and Echevarria (2011) on mathematics test scores; Bharadwaj et al. (2010) on mathematics test scores and attendance, for Chile.

[2] As one example of the temporal differences in neonatology, whereas 50 years ago the threshold for infant viability was around 1500 grams, today the threshold for viability is as low as 500 grams or even lower. As such, it is independently valuable to study the effects of birth weight using a more contemporary set of births than those used in the existing literature.

linking neonatal health to either educational or later outcomes in a western development context using very recent birth cohorts.[3]

We make use of a major new data source that can fill these gaps in the literature. We match all births in Florida from 1992 through 2002 to subsequent schooling records for those remaining in the state to attend public school. Florida is an excellent place to study these questions because it is large (its population of around 17 million compares to Norway, Denmark, and Sweden combined) and heterogeneous (45 percent of mothers are racial or ethnic minorities, and 23.4 percent of mothers were born outside the United States.) In addition, Florida is well-known for having some of the strongest education data systems in the United States; Florida, North Carolina, and Texas established the most advanced statewide student longitudinal data systems in the United States during the first half of the 2000s, and Florida has been testing children annually from third through eighth grade for over a decade. For several cohorts, Florida also implemented a universal kindergarten readiness assessment that allows us to explore the effects of birth weight on children's cognitive outcomes as early as age five. In addition to superb education data quality, Florida offers another major advantage when attempting to match birth and school records: Because children born in Florida are immediately assigned a social security number, and because social security numbers are collected upon school registration, Florida presents the opportunity for particularly effective matches between birth and school records. This allows Florida's health and education agencies the ability to nearly perfectly match births to school records. As we describe in the next section, our match rate is almost identical to what we would have expected based on American Community Survey data.

With these new data, we follow over 14,000 pairs of twins from birth through middle school to study the effects of birth weight on cognitive development. Ours is the first analysis of matched birth-school data in a developed context, and the first study to our knowledge conducted anywhere in the world that looks at heterogeneous effects across a variety of demographic and socio-economic dimensions. In addition, ours is the first study to explore the interaction between schooling factors and the relationship between birth weight and children's cognitive development.

We find that the effects of birth weight on cognitive development are roughly constant across a child's schooling career, and appear to be the same across a wide range of demographic and socio-economic groups. In addition, this trajectory is very similar regardless of the nature of the school the children attend. These results suggest that the gaps observed in adulthood associated with poor neonatal health are largely fixed at least by third grade or even kindergarten, indicating that some biological factors may be very difficult to overcome.

II. A new data source

A. Description of the data set and match diagnostics

We make use of matched data for all twins born in Florida between 1992 and 2002 and educated in a Florida public school afterward. For the purposes of this study, Florida's education and health agencies matched children along three dimensions: first and last names, date of birth, and social security number. Rather than conducting probabilistic matching, the match was conducted such that a child would be considered matched so long as (1) there were no more than two instances of modest inconsistencies (e.g., a last name of "Johnson" in the birth record but "Johnsen" or "Johnosn" in the school record, or a social security number ending in "4363" in the

[3] The potential benefits of using more current data become apparent when we compare the mean birth weight amongst twins in our study of children born after 1992 (2409 grams) to those from previous studies of twins born in the 1930s through the 1970s (which range from 2517 to 2598 grams, depending on the cohort and country.) Therefore, our population includes a set of twins who more closely resemble twins being born in western industrialized countries today.

birth record but ending in "4336" in the school record); and (2) there were no other children who could plausibly be matched using the same criteria. Common variables excluded from the match were used as checks of match quality. These checks confirm a very high and clean match rate: In the overall match on the entire (not just twin) population, the sex recorded on birth records disagreed with the sex recorded in school records in about one-one thousandth of one percent of cases, suggesting that these differences are almost surely due to typos in the birth or school records.

Between 1992 and 2002, 2,047,663 births were recorded by the Florida Bureau of Vital Statistics, including 22,625 pairs of twins. Of these children, 1,652,333 have been subsequently observed in Florida public school data maintained by the Florida Department of Education's Education Data Warehouse, and 17,639 pairs of twins have both twins present in the Department of Education data. All told, 79.6 percent of all children born in Florida, and 79.5 percent of twins born in Florida, are matched to school records using the match protocols.

In order to judge the quality of the match, we compare the 79.6 percent rate to population statistics from the American Community Surveys and Census of Population from 2000 through 2009. Recall that a child can only be matched in the Florida data if he or she (1) is born in Florida; (2) remains in the state of Florida until school age; (3) attends a Florida public school; and (4) is successfully matched between birth and school records using the protocol described above. Reasons (1) through (3) are "natural" reasons why we might lose children from the match. Our calculations from the American Community Survey indicate that, amongst the kindergarten-aged children found in that survey who were born in Florida, 80.9 percent were remaining in Florida at the time of kindergarten and were attending public school. This figure is an overstatement of the true expected match rate because the American Community Survey includes only children who are still living in the United States at the time of kindergarten. Given that some children born in Florida leave the country in their first five years because of emigration, because they were born to non-immigrant visitors to the country, or because they were born to undocumented immigrants who returned to their home countries, the true expected match is somewhat below 80.9 percent. While we are not aware of data regarding the fraction of U.S.-born children who leave the country before age five, it seems reasonable that the difference between 79.6 percent and 80.9 percent could be explained by these three reasons. We therefore conclude that the match rate is extremely high, and that nearly all potentially matchable children have been matched in our data.

B. Comparisons of the matched data set to the overall population

While we have demonstrated that we have matched virtually all potentially matchable children born in Florida to Florida school records, it is still the case that the set of Florida-born children attending Florida public schools differs fundamentally from the set of all Florida-born children. People who leave the state of Florida might differ from those who remain, and children attending public schools might differ fundamentally from those who attend private schools. It is therefore important to gauge how comparable the matched population is to the overall population of twins born in Florida. Though it is separate from how the matched data differ from the population, it is important to note that twins differ from singletons in important ways. We discuss issues of external validity in the conclusion.

Table 1 presents some evidence regarding the overall representativeness of our population of twins, along a number of dimensions: maternal race and ethnicity, maternal education, maternal age, maternal immigrant status, and parental marital status. There are four columns in the table: The first column reflects the total population of children born in Florida; the second column reflects the population of children matched to Florida public school records; the third column represents the set of children with a third grade test score; and the fourth column reflects the set of twins born in Florida who have a third grade test score. The comparison

between the first two columns demonstrates the total population of twins born in Florida; and the third column reflects the total population of twins born in Florida matched to Florida public school records. The comparison between the first and second columns makes clear the costs associated with carrying out this type of analysis in the United States, where children are lost for matching if they cross state lines between birth and school or if they attend private school. We observe that the set of matched children are more likely to be black (24.8 percent of matched children versus 22.6 percent of all children) and less likely to have married mothers (62.2 percent versus 64.8 percent of all children). The mothers of matched children are more likely to be less educated (17.5 percent college graduate versus 20.5 percent overall, and 22.5 percent high school dropout versus 20.9 percent overall) and are moderately younger (23.6 percent aged 21 or below versus 22.0 percent overall, and 9.4 percent aged 36 or above versus 9.8 percent overall). The comparison between the second and third columns of table 1 shows the difference in composition of the population of test-takers in elementary school versus those matched to school records more generally. As can be seen, 3rd-grade test-takers are still lower in terms of socio-economic status than are all children appearing in public school data. The fact that matched children are of somewhat lower socio-economic status, and that those with 3rd-grade scores are somewhat lower again, is unsurprising, given the well-documented relationship between family income (or parental education) and private school attendance.[4] However, our findings of estimated relationships between birth weight and test scores that are remarkably similar across very dissimilar groups reduces some of the potential concerns regarding external validity.

The comparison between the third and fourth columns of table 1 demonstrates the consequences of making use of twin comparisons, as is standard in the literature. As can be seen, mothers of twins are quite different from mothers of singletons: Mothers of twins are substantially less likely to be Hispanic or foreign-born and substantially more likely to be married than are mothers of singletons. In addition, they are considerably better educated (23.0 percent college graduate versus 15.8 percent in the overall population of test-takers, and 15.5 percent high school dropout versus 23.4 percent of all test-takers) and considerably older (13.6 percent aged 36 or above versus 9.1 percent in the overall population of test-takers, and 14.4 percent aged 21 or below versus 24.2 percent in the overall population of test-takers.)[5] Therefore, the decision to focus on twin comparisons to promote increased internal validity brings with it some cost in terms of external validity.

C. Birth weight distributions

1. Distribution of birth weight discordance

The variation that we use to identify the effect of poor neonatal health on cognitive skills comes from the fact that nearly all twin pairs differ in the birth weights of the two newborns, and sometimes the difference is quite substantial. In Florida, the average discordance in birth weight is 284 grams (0.63 pounds), or 11.8 percent of the average twin's birth weight of 2409 grams. Figure 1 presents the distribution of discordance for all twins, as well as all twins matched to test scores. As can be seen, the two distributions are virtually identical, so even though twins remaining in Florida and attending public schools have different maternal characteristics than do

[4] These relationships are observed in the Census data: In the 2000 Census, for instance, 6 percent of families earning $0-$25,000 per year sent their children to private school, as compared with 7 percent for those earning $25,000-$50,000 per year, 13 percent for those earning $50,000-$75,000 per year, and 19 percent for those earning over $75,000 per year.

[5] Twins are also more likely to be the consequence of in-vitro fertilization (IVF) or other forms of assisted reproductive technologies (ART). Later in this paper we investigate the differential effects of birth weight for twins likely conceived using IVF/ART versus those less likely to have been conceived using IVF/ART.

twins who leave Florida or attend private schools, the identifying variation does not differ at all. 51.3 percent of twin pairs have birth weight discordance over 200 grams, and 16.8 percent have birth weight discordance over 500 grams. 45 percent of twin pairs have birth weight discordance greater than 10 percent of the larger twin's birth weight, 26.6 percent have discordance greater than 15 percent of the larger twin's birth weight, and 14.7 percent have discordance greater than 20 percent of the larger twin's birth weight.[6]

2. Twins v. singletons

Figure 2 makes clear that twins have a dramatically different distribution of birth weight than do singletons. The mean twin birth weight during our time period is 2409 grams, 27.9 percent smaller than the mean singleton birth weight of 3342 grams. One can easily observe that for both twins and singletons the birth weight distribution of children observed in the test score data is identical to the distribution of all children born in Florida. 53.3 percent of twins have birth weights below 2500 grams (considered clinically low birth weight), as compared with 5.9 percent of singletons, while 7.1 percent of twins have birth weights below 1500 grams (considered clinically very low birth weight), as compared with 0.9 percent of singletons.

Note that the average birth weight in our population is considerably smaller than those in other studies using children from developed western countries born a generation or two earlier: The mean birth weight in Behrman and Rosenzweig's (2004) study of Minnesota twins born 1936-1955 was 2557 grams; for Royer's (2009) California twins born 1960-1982, Black, Devereux, and Salvanes's (2007) Norway twins born 1967-1981, and Oreopoulos et al.'s (2009) Manitoba twins born 1978-1985, the mean birth weights were 2533 grams, 2598 grams, and 2517 grams, respectively. The lower mean birth weight in our sample is almost surely the result of improvements in medical technology that allow lower birth weight babies to survive longer. This change in technology and shift in the distribution of birth weights highlights another benefit of studying children born in the 1990s and 2000s.[7]

III. Empirical framework

Our empirical framework largely follows what has become standard in the literature. We estimate twin fixed effect models in which the regressor of interest is the natural logarithm of birth weight.[8] Following Almond, Chay and Lee (ACL, 2005) and Black, Devereaux and Salvanes (BDS, 2007), let

$$y_{ijk} = \alpha + \beta \ln(bw)_{ijk} + x'_{jk}\gamma + \phi_{jk} + \varepsilon_{ijk} \quad (1)$$

where i indexes individuals, j indexes mothers, k indexes births, y_{ijk} denotes the outcome of child i, born to mother j in twin-pair k, x is a vector of mother and child-specific determinants of the

[6] There exists medical evidence that large birth weight discordances lead to increased chances of severe disability. For instance, Luu and Vohr (2009) find that the likelihood of cerebral palsy in a twin is four times greater when birth weight discordance is over 30 percent than when it is less than 30 percent.
[7] Note also that the rate of the twinning in the developed world increased 76 percent between 1980 and the present (Ananth and Chauhan, 2012).
[8] We estimate other model specifications as well to explore the degree to which our results are robust to model specification.

outcome, ϕ denotes unobservable determinants of the outcome that are specific to the mother and birth, and ε is an error term.[9]

For the majority of specifications presented in the paper the outcome, denoted *y*, is a test score – the criterion-referenced Florida Comprehensive Assessment Test (FCAT) – which is standardized within grade and year to have mean zero and standard deviation one in the entire population of children in Florida.[10] (Our measures of kindergarten readiness are dichotomous, so we estimate linear probability models in those cases.) For ease of presentation, we average standardized reading and mathematics FCAT scores for our dependent variable, but our results are presented separately for reading and mathematics, and the test-specific results are available on request. The regressor of interest, ln(*bw*), is the natural logarithm of birth weight in grams. In section 6 we present results from specifications other than the linear-in-log model, but the linear-in-log model appears to fit the data well.

Ordinary Least Squares (OLS) estimation of (1) would produce biased estimates of β if ϕ_{jk} were correlated with $\ln(bw)_{ijk}$ – in other words, if there were unobservable determinants of cognitive ability that were correlated with birth weight. To address the potential bias due to correlation between ϕ_{jk} and $\ln(bw)_{ijk}$, we estimate a twin fixed effect model.

Twins necessarily share the same x_{jk} and ϕ_{jk}. A twin fixed effect model essentially differences out any mother- or birth-specific confounder and identifies β based on between-twin variation in test scores and birth weight. Logically, birth weight can vary due either to variation in gestation length, or to variation in fetal growth rates. By focusing on twins, we necessarily hold gestation length constant. Our estimates are identified, therefore, by variation in fetal growth rates.

One potential internal validity concern is that we can only make use of test score data for a twin pair if both members of the pair have test scores. If one twin is present in the test score data but not the other, and the reasons for differential inclusion in the data are correlated with neonatal health, this could present a source of bias. There are three different reasons why we might observe differential inclusion in the test score data related to poor neonatal health. First, parents may send one child to public school but the other to private school; if parents systematically send their heavier or lighter twin to different schooling settings, this could affect the observed relationship between birth weight and test scores, conditional on being in the public school setting. Second, since Florida exempts students from the FCAT in case of severe disability,[11] any relationship between birth weight and the likelihood of severe disability could affect our estimated relationship of interest. A third potential reason is similar to the second: If low birth weight children are more likely to miss the exam because of illness or absence, the effect on the estimates would be similar to the bias that results from differential disability.

That said, the evidence suggests that these potential internal validity concerns are not major issues. When we estimate twin fixed effect regression models in which the dependent variable is whether the twin ever attended public school (79.5 percent of the Florida birth population) the coefficient on log birth weight is -0.012 (with a standard error of 0.008). If the dependent variable is an indicator for beginning in public school by first grade (77.0 percent of all

[9] In practice, because we estimate models with twin-pair fixed effects, the mother-specific determinants of the outcome are subsumed in the fixed effect. The individual-specific determinants are child gender and within-twin-pair birth order.

[10] We standardize FCAT scores for ease of interpretation. Our results are not substantively changed if instead we measure the FCAT in its unstandardized developmental scale score format.

[11] Florida's Final Rule 6A-1.0943 gives the grounds for FCAT exemption, stating students can be exempted from the test in "extraordinary circumstances [that] are physical conditions that affect a student's ability to communicate in modes deemed acceptable for statewide assessments."

births), the coefficient estimate is -0.006 (with a standard error of 0.008). If the dependent variable is whether we ever observe an FCAT score for the child (69.1 percent of all births), the coefficient estimate is -0.003 (standard error of 0.009), and if the dependent variable is whether we observe an FCAT score in every possible year when we would have expected to see one if the student did not leave Florida public schools (65.5 percent of all births) the coefficient estimate is 0.006 (standard error of 0.004). In sum, it appears that relatively heavy and relatively light twins are remaining in public school and taking the FCAT at highly similar rates. This diminishes the potential internal validity concerns associated with differential test-taking rates.

IV. Preliminary results – heavier versus lighter twins

A. Heavier v. lighter twins

Before presenting the main regression results, we begin with simple comparisons of the test scores of heavier and lighter twins based on birth weight. These results, which aggregate twin pairs with small birth weight discordance with those with large birth weight discordance, are shown in figures 3 through 5.[12] Nonetheless, they clearly demonstrate the first main result of the paper. Figure 3 shows the average within twin pair difference in test score between the higher birth weight twin and the lower birth weight twin, while figures 4 and 5 show the same patterns for mathematics and reading, respectively. These figures show test score differences for the average of math and reading scores calculated separately at grades three through eight, along with the 95-percent confidence interval around those differences.[13][14] Note that these figures do not reflect panels of students, so there are different groups of children in each grade.

Within twin pairs, on average the heavier born scores about five percent of a standard deviation higher than his lighter born twin. This difference in test scores is statistically distinguishable from zero, and is stable from third through eighth grades, covering ages from approximately 9 to 14. This comparison holds constant any confounding factor that varies at the family, mother or birth level. The results imply that fetal health, as proxied by birth weight, has effects on cognitive skills by age 9. Furthermore, this effect does not seem to either dissipate or widen through preadolescence.

B. Testing for differential attrition

As described above, we observe test scores only for students who are in Florida public schools, and a small fraction of public school students miss the exam because of absence or

[12] We conduct this aggregation in order to fix ideas, but it is important to remember that birth weight discordance is known to be an independent predictor of major morbidities such as congenital anomalies, low APGAR scores, and periventricular leukomalacia (Vergani et al, 2004; Cheung, Bocking, and Dasilva, 1995).

[13] Throughout the analysis, unless otherwise noted, test score results are for the average of math and reading scores for observations with non-missing scores for both tests. For observations with one test missing, the non-missing test is used. In the main regression specification, 99.5 percent of observations have both math and reading scores, 0.2 percent have only a math score and 0.3 percent have only a reading score.

[14] For all analyses separated by grade, we assign students to the grade they would have been in had they progressed one grade per year from the first time we observe them with an FCAT score in third grade We use this "imputed grade" rather than the student's actual grade because grade retention may be affected by birth weight and because we are interested in following children longitudinally. All results are extremely similar if we focus on actual grade rather than this imputed grade.

because they have a profound disability.[15] Attrition from the testing data is only a concern for our estimates if students with missing test scores would have had particularly high or low scores relative to their twin, and since we are including twin-pair fixed effects in all of our models, attrition only causes bias if one twin leaves the testing data and the other remains. As mentioned above, it may be the case that parents might differentially send either heavier or lighter twins to private school, and it may also be the case that twins with poor neonatal health are more likely to be sickly and more prone to miss the FCAT, or more likely to be severely disabled and exempt from taking the FCAT.[16]

In addition to the twin fixed effects estimates described above, we present three tests of whether non-random attrition from the sample biases our estimates. Each of these tests indicates that the estimates are not meaningfully biased by non-random attrition. First, in figure 6 we plot the difference in test scores between heavier and lighter birth weight twins, restricting the sample to a sample for which we observe both twins for each of the six grades. The pattern is essentially unchanged from what we saw for the full sample in figure 3. The stability of the difference in test scores between heavier and lighter birth weight twins does not appear to be affected by changing selection out of the sample as twins age.

Second, we measure directly the amount of differential attrition from the sample between third and eighth grade. Starting with the sample of twins we first observe in third grade, figure 7 shows the difference in the fraction of heavier and lighter birth weight twins tested in each subsequent grade. The figure shows that lighter birth weight twins are slightly more likely to have missing tests in the sixth through eighth grades, possibly because they are pulled from public schools and possibly because they are still enrolled in public school but missed the exam.[17] However, the magnitude of the difference is not large enough to significantly affect the relative magnitude of the within twin pair test score differences, a conjecture also we test directly below.

Third, we can put bounds on the magnitude of any bias resulting from differential attrition of high and low birth weight twins between third and eighth grade. Figure 8 shows two sets of estimated differences by grade, one where we replaced missing test scores with the 5th percentile of test scores in that grade and another where we replaced missing test scores with the 95th percentile. As the figure shows, even assuming that students who no longer had an FCAT score after third grade had extremely low or extremely high test scores does not change the conclusion that the within twin pair difference in test scores is remarkably stable from third through eighth grade. Taken together, the results show that attrition patterns do not significantly affect the patterns of results between third and eighth grade.

V. Main results

We now turn to our main regression results. As described above, the basic regression model is an ordinary least squares estimate that includes twin-pair fixed effects, a gender dummy, and a dummy for within-twin-pair birth order. The dependent variable is the standardized FCAT

[15] While over 30 percent of Florida twins receive some special education services, a large majority of students with disabilities in Florida take the FCAT. Only students with disabilities such as severe mental retardation or severe autism are exempt from the FCAT.

[16] There has been a secular trend toward more very low birth weight infants surviving and entering the educational system impaired. See Zwicker and Harris (2008) and Aarboudse-Moens et al (2009) for examples of evidence.

[17] We start off with a sample of twin pairs with twins old enough for 3rd grade. We have 2.9, 2.1, 2.1, 2.1, 1.9 and 1.7 percent of pairs in grades 3, 4, 5, 6, 7 and 8 respectively where twins are old enough to be in the grade but neither of them is tested. We have 2.4, 2.3, 2.3, 2.4, 2.3 and 2.2 percent of pairs in grades 3, 4, 5, 6, 7 and 8 respectively where twins are old enough for the grade but we observe test scores for only one twin. We have already demonstrated above that in cases in which only one twin is missing a test score, it is not systematically the case that the twin with the missing score is the lighter twin.

score, and the regressor of interest is the natural logarithm of birth weight in grams. We report some results based on separate regressions for each grade from three through eight, and other results that pool test scores across all six grades. In the pooled regressions, standard errors are clustered at the individual level to account for the fact that each individual has up to six observations, one for each grade in which he or she was tested.

The non-parametric plots of the relationship between test scores and birth weight reported in figure 9 present evidence supportive of the log birth weight specification that we employ, as the figure is consistent with a concave relationship between birth weight and test scores. The figure shows two series, each derived from a test score regression that pools 3-8 grades and both math and reading scores. Each series plots the coefficients from a set of 36 dummy variables corresponding to 100g-wide birth weight bins. The bins range from a low of 501-600g to a high of 4,001-4,100g. In both regressions, the left-out group is below 501 grams. As was observed in similar sets of plots by ACL and BDS, the figure also shows clearly that the shape of the relationship between test scores and birth weight is very similar both unconditional and conditional on twin-pair fixed effects.

A. Pooled results for full sample

We present our main result in column 2 of table 2. The results show that within twin-pairs higher birth weight is indeed associated with higher test scores in grades three through eight. The estimated coefficient of 0.441 implies that a ten percent increase in birth weight is associated with a 0.045 standard deviation increase in test scores. The coefficient is precisely estimated, with a *t*-statistic of over 15. The fixed effects result is modestly larger than, but in the same general ballpark as, the equivalent OLS coefficient of 0.310 reported in the first column of table 2. The results are somewhat larger for mathematics than for reading, but the patterns are the same for both subject areas; therefore, for ease of presentation, we concentrate on the combined mathematics and reading results for the remainder of the paper.

To put the magnitude into perspective, BDS estimate that the effect of log birth weight on log earnings is 0.12. Assuming the log wage return to cognitive skills is 0.2 as estimated by Neal and Johnson (1996), our estimates imply that increases in cognitive skills present in grades three through eight explain approximately three-quarters of the effect of birth weight on wages found by BDS. Our estimate of the effect of neonatal health on cognitive development is large in those terms, but it is worth comparing to other important correlates of student achievement. Figure 10 shows the test scores of heavier and lighter born twins stratified by mother's education. The figure clearly shows that the difference in test scores resulting from differences in birth weight is small compared with differences in achievement associated with mother's education. Each of the differences between heavier and lighter born twins shown in the figure is statistically significant. However, it is clear that in terms of math and reading achievement, it is better to be the lighter born twin from a college educated mother than the heavier born twin from a high school dropout mother. Taken together, these findings suggest that while "nurture" can go a long way toward remediating a child's initial disadvantage, there are still biological factors at play that make it difficult to fully remediate this disadvantage.[18]

B. Results by grade for full sample

A key question of interest is how the cognitive effects of *in utero* conditions and neonatal health develop. We have already shown that the effects of birth weight on cognitive achievement

[18] By this statement we do not mean to suggest that the results answer the age-old nature-nurture question. Rather, they are consistent with the growing literature on epigenetics that shows that environmental and biological factors interact (Miller et al., 2009 or Lam et al., 2012)

in grades three through eight are similar to those observed with respect to adult earnings. We next explore how the impact on test scores changes during these important years for human capital development. Does the effect of birth weight grow larger as children age, or is it present by age 9 and does it remain constant through the upper elementary and middle school grades? The structure of the data allows us to estimate the effect of birth weight on test scores separately by grade to address these questions.

The results are presented in columns 3-8 in table 2. The table shows the estimated effect of log birth weight from twin fixed effects models that are estimated separately for test scores from each grade, three through eight. As is the case throughout the paper, grade refers to grade an individual would have attended if he had progressed on a normal schedule after we first observe him or her take a third grade exam. (We call this the "imputed grade.") We have estimated the models based on actual grade, by dropping all twin pairs that do not progress on a normal schedule, and by age rather than grade. The results using these alternative specifications are substantially the same as the ones we present.

The table shows that the effect of birth weight on cognitive achievement is in fact present by the third grade. The twin fixed effect estimate of the effect of log birth weight on test scores in third grade is 0.442. The grade-specific estimated effect remains fairly stable from third through eighth grade, ranging from 0.373 to 0.526. Note that while the *F*-test that the grade-level estimated effects are identical is rejected at a moderate level of statistical significance (p=0.057), there is no evidence that this follows a substantial systematic pattern as children age. In a regression model in which we interact the log of birth weight linearly with grade in school, the coefficient estimate on the interaction term is one-two thousandth the magnitude of the coefficient on the log of birth weight. These results suggest that the effects of neonatal health do not substantially change, or develop, between ages 9 and 14. Rather, whatever effect health at birth has on cognitive development occurs largely by age 9, and remains fairly constant throughout the preadolescent and adolescent years. Furthermore, comparing the effect with the results BDS find for adult earnings suggests that the effect of birth weight appears not to change substantially between 14 years and prime working age.

C. Role of genetic differences between twins

For some policy conclusions we might draw from the results, it could be important to isolate the impact of factors that change intrauterine growth while holding genetics constant. In studies where the zygosity of twins is known, it is possible to restrict attention to comparisons between monozygotic twins, effectively holding genes constant.[19,20] A potential weakness of our data is that they do not include the zygosity of the twins. We do, however, know the gender of each child, and can use this information to obtain some purchase on whether the relationship between birth weight and test scores is driven by within-twin pair differences in genetics. Same-sex twin pairs are a mix of monozygotic and dizygotic pairs. Different-sex twin pairs are, however, all dizygotic. If genetic differences were driving a significant portion of the relationship between birth weight and test scores, and birth weight were positively correlated with positive determinants of later cognitive skills, we would expect to see a stronger correlation between birth

[19] Zygosity refers to whether the twins are identical (monozygotic) or fraternal (dizygotic). Monozygotic twins form from a single egg, or zygote, and split into two embryos. Dizygotic twins form from two separate eggs, fertilized by different sperm. Monozygotic twins therefore share the same genes, whereas dizygotic twins generally have the same genetic similarity as two full siblings.

[20] Research on epigenetics implies that while they do hold constant genetics, comparisons among monozygotic twins do not hold constant all of the effects emanating from genetics. Epigenetics is the study of the way that environmental factors interact with genes to influence which genes are expressed.

weight and test scores among different-sex twin pairs. The first panel of table 3 shows estimates separately for same-sex and different-sex twins.

Note that as we explore the heterogeneity in effects across different groups, we restrict our attention to models that include all test scores from grades three through eight, restricting the effect of birth weight to be the same at each grade. For the sake of clarity, in our main tables we report the results in which we pool test scores across all grades; in Appendix table A1 we report grade-by-grade results for all subgroups. The first column of table 3 reports the percent of the population in each subgroup. The second column reports the mean test score for the group. The third column reports the mean and standard deviation of the group's birth weight. The fourth column reports the mean and standard error of the estimated effect of birth weight on pooled third through eighth grade test scores. The last column reports the *p*-value from an *F*-test of the null hypothesis that the estimated birth weight effects are the same across relevant groups.

Turning to the results, the estimated effect of birth weight is virtually identical for same-sex twins (0.447) and different-sex twins (0.427), suggesting that the estimated relationship is within the same general range regardless of zygosity. This result is consistent with results reported in BDS, who find no significant differences in the effect of birth weight on adult earnings between same-sex and opposite-sex twins. BDS also find no significant difference in estimated effect of birth weight on earnings for monozygotic twins and dizygotic same-sex twins in their sample with available zygosity information. Taken together, the results suggest that genetic differences between twins are unlikely to be driving a large portion of the relationship between birth weight and later life outcomes.

D. Differences by child gender

We next turn to an examination of how the effects of neonatal health vary across the population. We begin with a comparison by gender. This comparison allows us to examine the basic question whether the effects of birth weight on test scores are different for boys and girls (Rosenzweig and Zhang, 2009). With these results we begin to tell a story about consistency in the effects across the population that we will continue to explore in subsequent sections.

The results broken down by gender are shown in the second and third panels of table 3; the second panel includes girls and boys from both same-sex and opposite-sex twin pairs, while the third panel restricts the analysis to girl-girl and boy-boy pairs. Note that the mean test scores of boys and girls in the twins sample are 0.048 and 0.099, and boys are born 4.4 percent heavier (2473 versus 2369 grams). However, we cannot reject the hypothesis that the marginal effect of birth weight is the same for boys and girls. The estimate among female-female twin pairs is 0.444, and the estimate among male-male twin pairs is 0.449. These differences are not statistically distinguishable. The same patterns are seen when looking at all boys or all girls regardless of the gender composition of the twin pair. This is unsurprising: Recall that the estimated effects of birth weight are statistically indistinguishable between same (0.447) and opposite (0.427) sex pairs.

E. Results by maternal race, ethnicity and immigrant status

Two special features of the Florida context allow us to investigate heterogeneity in the effects of birth weight in ways that have not been possible in other related work to this point. Florida has a remarkably heterogeneous population. Approximately one-quarter of all births in Florida are to black mothers, 18 percent of births are to Hispanic mothers, and 18 percent to foreign-born mothers. The diversity of demographics in the state, combined with the size of the dataset make comparisons of birth weight effects across racial and ethnic groups possible.

It is inherently interesting to learn about whether the long-term effects of *in utero* conditions on cognitive development vary across demographic groups. Beyond this inherent

interest, examining heterogeneity in the effects may shed light on the mechanisms by which neonatal health affects cognitive skills. There are significant differences in household income, wealth and education by race, ethnicity and immigrant status. If these factors, each of which is strongly correlated with student achievement at the population level, are substitutes with neonatal health in the production of cognitive skills then we should expect to see larger effects of birth weight on test scores for more disadvantaged groups. If income, wealth and parental education are complements with neonatal health, one would expect to see larger effects for more advantaged groups.

The fourth through sixth panels of table 3 shows estimates of the effect of birth weight on pooled third through eighth grade test scores separately by race (panel 4), ethnicity (panel 5) and immigrant status (panel 6). Births to black mothers account for 26.1 percent of the sample, while births to white mothers account for 72 percent. Twins with black mothers score 0.722 standard deviations lower than twins with white mothers, a gap that is of a similar magnitude with black-white test score gaps measured in the same time period for random samples of U.S. school children (Chay, Guryan, and Mazumder, 2009). The gap in average test scores between students with a Hispanic and non-Hispanic mother is smaller but still substantial, 0.459 standard deviations. There is virtually no difference in test scores between the children of immigrant and non-immigrant mothers.

Despite the fact that there are substantial differences in average test scores between demographic groups, we estimate no statistically or economically significant differences in the effect of birth weight on test scores across these groups. Furthermore, there is no clear pattern in the point estimates. The estimated effect of birth weight on third through eighth grade test scores is somewhat smaller for twins with black mothers than for twins with white mothers (0.381 versus 0.466). However, the estimated effect is somewhat larger for twins with Hispanic mothers than for twins with non-Hispanic mothers (0.478 versus 0.434), though not statistically significantly so. And the estimated effects for twins with immigrant and non-immigrant mothers are quite close (0.449 versus 0.440) and statistically indistinguishable from one another. Taken together, these results suggest that the effect of birth weight on cognitive development is remarkably consistent across demographic groups.

F. Results by family socio-economic status

Based on these results, there does not appear to be a systematic relationship between the effect of birth weight and demographic characteristics that are related to parental income and human capital. The data also allow us to test for these relationships more directly. The seventh and eighth panels of table 3 show effects estimated separately by mother's education (panel 7) measured at the time of the birth and a proxy of family income (panel 8) – the median income in the zip code of residence at birth. Sixteen percent of the population of twins with test scores are born to high school dropout mothers, 61 percent of the sample are born to mothers with a high school degree and/or some college, and 23 percent of the sample are born to college graduate mothers.

There is a strong relationship between mother's education and children's test scores. On average in third through eighth grade, children with college graduate mothers score more than a full standard deviation higher than children with high school dropout mothers, and two-thirds of a standard deviation more than those with high school graduate mothers. High school and college graduate mothers also have slightly higher birth weight twins than high school dropout mothers.

Estimated birth weight effects also vary somewhat by mother's education. The estimated effects of log birth weight on pooled third through eighth grade test scores are monotonically increasing across the three categories of mother's education. For twins with high school dropout mothers the estimated coefficient on log birth weight is 0.359; for those with a high school graduate mother the estimated coefficient is 0.434; for those with a college graduate mother it is

0.529. None of these estimates is statistically distinguishable from any other, and an *F*-test fails to reject the hypothesis that the coefficient is the same for all three maternal education groups. It is thus important to be careful about the inferences we draw from these differences. That said, it is worth noting that one might have expected the effects of birth weight to be weaker for better-educated families than for worse-educated families, rather than the reverse, if family inputs were substitutes for neonatal health. Indeed, when we estimate a model with both log birth weight and the interaction between log birth weight and maternal years of schooling, we find an estimate of the interaction term of 0.028 with a standard error of 0.012. The fact that we observe strong relationships between birth weight and test scores across all maternal education groups strengthens the notion that while some biological disadvantage can be overcome, there remain some biological factors that are very difficult to overcome with nurture.

This conclusion is reinforced by a split by a proxy for family income – the median income (as of the 2000 Census) in the zip code of residence at the time of birth. As seen in panel 8 of table 3, there is a strong relationship between test scores and this measure of family resources: Those children whose families resided in the one-third richest zip codes at the time of birth score two thirds of a standard deviation higher on tests than did those who resided in the one-third poorest zip codes at the time of birth.[21] As in the case of maternal education, we do not observe a statistically significant difference across birth neighborhood affluence groups in the estimated effects of birth weight on children's test scores, though the point estimate is modestly larger for the two more affluent groups than for the least affluent group.[22] This provides one more piece of evidence that family resources might help to mitigate biological factors but they are very difficult to completely offset.

We can also explore differences by maternal marital status at the time of birth (panel 9) and maternal age at the time of the twins' birth (panel 10), both reported in table 3. Approximately two-thirds of the twins in our sample are born to married mothers. While birth weights of twins born to married mothers are only slightly higher than those born to unmarried mothers, there is a large difference in test scores; twins born to married mothers have average test scores that are nearly two-thirds of a standard deviation higher than those born to unmarried mothers. There is suggestive evidence that the effect of birth weight may be larger among married mothers than unmarried mothers. The point estimate on log birth weight is 0.485 for married mothers and 0.362 for unmarried mothers. This is the one head-to-head comparison that is statistically significant at conventional levels – the *p*-value of the difference is 0.064 – though the magnitude of the results are qualitatively quite similar, suggesting that the effects of poor neonatal health on cognitive outcomes are of approximately the same magnitude across a wide range of demographic and socio-economic dimensions. But to the extent to which the estimates are larger for children of married mothers than for children of unmarried mothers, this is the opposite of what one might have expected to find if biological differences were not as important in families with greater parental resources. While higher-socio-economic-status families clearly remediate early health disadvantage to a great degree (witness the fact that poor-health children in educated families perform much better than good-health children in less educated families) there seems to be a portion of this disadvantage that is persistent and much more difficult to remediate, and if anything, the evidence points toward neonatal health and parental inputs being more likely to be complements than substitutes. This point, of course, is necessarily somewhat speculative.

We separate maternal age into four categories: less than or equal to 21, 22-29, 30-35 and older than 35. Fifteen percent of twins in the sample were born to mothers who were 21 years or

[21] Note that the sizes of these groups differ by the time children are in school because, as noted above, more affluent families are more likely to send their children to private school.

[22] When we estimate a model with both log birth weight and log birth weight interacted with zip code median income (in $1000s), the coefficient estimate on the interaction term is -0.0004 with a standard error of 0.0003.

younger, 40 percent were born to mothers between 22 and 29 years old, 32 percent were born to mothers who were 30-35, and 14 percent were born to mothers older than 35. There is a strong positive relationship between average third through eighth grade test scores and mother's age at birth. Twins born to mothers older than 35 have test scores that are almost three-quarters of a standard deviation higher than those born to mothers 21 years old and younger, though the nature of twinning indicates that some caution should be taken in interpreting this correlation. Selection into twinning differs with mother's age because of physiological changes and differences in use of fertility treatments. Despite the strong relationship between mother's age and average test scores, there is no relationship between mother's age and the estimated effect of birth weight. Estimates of the effect of log birth weight on pooled third through eighth grade test scores range from 0.372 to 0.483 among the four mother's age categories. The relationship between point estimates and mother's age is non-monotonic across these categories and the *p*-value of the *F*-test of the hypothesis that they are all equal is 0.698.

G. Summarizing heterogeneity in birth weight effects on cognitive development

Our general conclusion after considering these ten different dimensions over which the effects of birth weight on test scores is that these effects are roughly the same for children from a wide range of different backgrounds – evidence that the effects of biological factors are present throughout the socio-economic distribution. We note, however, that if anything it appears that relatively high-socio-economic status families experience larger, rather than smaller, effects of birth weight, suggesting that neonatal health and family inputs may be complements rather than substitutes. Figure 11 places this possibility into visual focus: We plot all 18 subgroups' point estimates against the mean test scores for that group – a range greater than a full individual-level standard deviation of the test score distribution.

The figure demonstrates two important features of the heterogeneity of birth weight effects across demographic groups. First, the estimated effects of birth weight are all within the same general range between 0.36 and 0.53, and the estimated effects are both statistically and economically significant for every demographic and socio-economic group we analyzed. A log birth weight effect of 0.36 would indicate that effects on cognitive development could account for 60 percent of the long-term relationship between birth weight and test scores estimated by BDS. At the other end of our range, an estimate of 0.53 would indicate that effects on cognitive development could account for 88 percent of the BDS estimate.

The second pattern the figure demonstrates is that there does appear to be an upward-sloping relationship between estimated treatment effects and the subgroup's mean test score. This positive relationship indicates that the effects of birth weight are somewhat larger for relatively advantaged groups of children than they are for relatively disadvantaged groups of children. The slope of the line plotted in figure 11 is 0.125, with a standard error of 0.019, and is highly statistically significant. While by no means definitive, this pattern indicates that biological factors may modestly disproportionately inhibit high socio-economic status families, and is suggestive that neonatal health and parental resources are complementary.[23]

[23] Children in higher-scoring subgroups – such as those from high income, highly educated families with older mothers – are more likely to have been born with the assistance of in-vitro fertilization (IVF) or other assisted reproduction technologies (ART). It is therefore conceivable that the positive relationship plotted in figure 11 is due at least in part to differential patterns of IVF/ART. This could be especially important in a population of twins, given that Bitler (2008) demonstrates that IVF/ART substantially increases the likelihood that a mother will have twins, and that twins likely conceived with the assistance of IVF/ART have lower-quality birth outcomes. While we cannot measure IVF/ART use in our data, we conduct two checks to see whether or not differential IVF/ART prevalence is a plausible explanation for our findings. First, we conduct the identical analysis for twins born to mothers aged 30 and above, versus those under 30;

VI. Effect variation across the birth weight distribution and with birth weight discordance

Thus far, we have presented estimates of our baseline model, which specifies that the relationship between average test scores and birth weight is linear in the log of birth weight. Understanding how the marginal effect of birth weight varies across the birth weight distribution and with birth weight discordance may be helpful in narrowing down potential mechanisms for the relationship. There is great attention paid by public health officials and medical practitioners on the thresholds of 1500g and 2500g, the conventional delimiters of very low birth weight and low birth weight. Stronger marginal effects of proportional increases in birth weight for very low and low birth weight babies might suggest different physiological mechanisms than if the effect were only present in comparisons between moderate and high birth weight babies.

We have already presented non-parametric evidence (figure 9) that the relationship between birth weight and student test scores appears to be concave, supporting the log birth weight specification that is common in the related literature. That said, there could still be some important nonlinearities in the relationship. In this subsection we relax the assumptions underlying our main specification in two additional ways, and in doing so explore how the marginal effect of poor neonatal health varies across the distribution of birth weight and with birth weight discordance. For one, we estimate models that allow the marginal effect of birth weight to vary across different bins of the birth weight distribution. This analysis includes models that are fairly non-parametric in the specification of these marginal effects, and models that test explicitly whether the marginal effects are different for very low birth weight (<1500g), low birth weight (1500-2499g) and normal birth weight (>2500g) babies. We also include models that allow the effect of neonatal health to vary nonlinearly according to the discordance in birth weight between twins. In addition, we investigate whether alternative parametric functions of birth weight better capture the relationship of interest. For example, using slightly more parametric models, we test whether the marginal effect of an additional gram of birth weight is constant, or if the marginal effect is constant in percentages.

In figure 12, we present estimates from a regression specification that address the former of these questions. The estimates come from a regression that is based on our standard twin fixed effect specification. The only difference is that log birth weight is interacted with a set of dummy variables corresponding to 20 bins, each of which corresponds to 5 percent of the lighter-born twin's birth weight distribution. These interactions allow the marginal effect of log birth weight to vary freely across the bins. We have also estimated models that define the bins based on the heavier-born twin's birth weight. These results are very similar and are presented in appendix figure A1. The results show no systematic relationship between the marginal effect of birth weight on test scores and the level of birth weight. The estimated effects are largely stable, aside from variation that appears to be due to sampling variation, across the distribution of birth weight.

---

this is the age breakdown that Bitler uses to proxy for IVF/ART likelihood. The estimated slope of the line for the 30-and-up group is 0.127 (standard error of 0.045) while the estimated slope of the line for the under-30 group is 0.114 (standard error of 0.042); the p-value of the difference between these two slopes is 0.842. Next, we conduct the identical analysis for twins who were the first children born to the mother to those who were not the first children born to the mother, given that IVF/ART is more likely amongst families with previous fertility challenges. The estimated slope of the line for the first-children mothers is 0.067 (standard error of 0.038) and the estimated slope of the line for the subsequent-pregnancy mothers is 0.152 (standard error of 0.031). While the difference between these two slopes is modestly statistically significant – the p-value is 0.091 – the positive relationship between birth weight effects and SES is stronger for the group of twins *less* likely to be conceived via IVF/ART. Taken together, these results suggest that differential probabilities that children from high-scoring subgroups were conceived via IVF/ART are not responsible for the positive-sloped relationship between the scoring level of the subgroup and the subgroup-specific estimated effect of birth weight on test scores.

There appears to be somewhat more variation in the estimated effects at higher birth weights, but an F-test fails to reject the null hypothesis that the coefficient on log birth weight is the same across all 20 bins (p-value: 0.840).

We explore the second of these questions – whether the relationship between birth weight and test scores varies by birth weight discordance – in figure 13. We divide twins into 20 bins by birth weight discordance, excluding the twin pairs that are very close in weight (<150g difference). As can be seen in the figure, the estimated relationship between birth weight and test scores is qualitatively similar across a wide variety of birth weight discordance.

The non-parametric results presented in figure 9 suggest the marginal effect of log birth weight on test scores is fairly stable across the birth weight distribution. But the salience of 1500g and 2500g, both among medical professionals and in the social science literature on early life health, lead us to estimate specifications that test whether the marginal effect of birth weight on test scores varies above and below these thresholds. Therefore, in rows 2 and 3 of table 4 we present results from specifications that allow the effect of log birth weight to be different above and below 2500g. To do this, we estimate one effect for twin pairs where both twins were born heavier than 2500g, and another where both twins were born lighter than 2500g. In both cases, we estimate the baseline specification in which the effect of log birth weight is assumed to be constant within the group.

The estimated effect of a marginal increase in birth weight is quite similar for low birth weight (<2500g) and normal birth weight (≥2500g) children. The estimate for low birth weight twin pairs is 0.473, and for normal birth weight twin pairs it is 0.526. The two pooled coefficients are not statistically distinguishable (p-value: 0.535). The results reported in rows 4 and 5 of the table estimate different log birth weight effects for two additional groups: very low birth weight (<1500g) and low birth weight (1500-2499g). Consistent with the previous results, the estimated effects do not vary significantly across these groups. The estimated effects for very low birth weight, low birth weight and normal weight are, respectively, 0.572, 0.517 and 0.526. An *F*-test fails to reject that these three estimates are the same (p-value: 0.914).

We also seek to more formally test the assumption suggested by our non-parametric estimates that the linear in log birth weight specification is reasonable. Rows 6 and 7 of table 4 show results from models that replace the natural logarithm of birth weight in equation (1) with two alternative specifications. The sixth row reports the result from a regression that replaces the log of birth weight with birth weight in thousands of grams, but which is otherwise equivalent to the baseline specification. When we restrict it to have a constant linear effect, we estimate that a marginal increase of 1000g of birth weight is associated with 0.186 standard deviations higher third through eighth grade test scores. The estimated effect is strongly statistically significant. As was the case in the log birth weight specifications, estimates that allow the effect to vary by grade are largely stable between third and eighth grade.

To test whether the linear-in-grams model fit the data we also estimated a model that allowed the marginal effect of a gram to be different among heavier and lighter twin pairs. Specifically, we interacted birth weight in grams with the average of the twin pair's birth weight, a specification reported in row 7 of table 4 (there are two coefficients reported in this row, the coefficient on birth weight and the coefficient on its interaction with the deviation from mean twin pair birth weight.)[24] The results show that the marginal effect of a gram of birth weight is smaller in heavier twin pairs, as indicated by the negative and significant coefficient on the interaction term. This result is consistent with the linear-in-logs model, in which test scores are proportionally related to birth weight. Based on these results, we concluded that the linear-in-logs

[24] The twin pair average birth weight is demeaned by the sample average so that the birth weight coefficient represents the marginal effect of birth weight in a twin pair of average birth weight. The main effect of twin pair average birth weight is subsumed by the twin pair fixed effects.

specification is a good approximation of the relationship between birth weight and cognitive skills.[25]

VII. School quality and the effect of birth weight on test scores

The results presented thus far have demonstrated that there is a robust relationship between birth weight and third through eighth grade test scores, and that this relationship is remarkably stable as children age through preadolescence, across different demographic groups, and across different socio-economic groups. The stability of this relationship is all the more notable because the marginal effect of birth weight does not vary very much across groups that have very different average test scores. Children growing up in circumstances that lead to very different achievement levels nonetheless appear to be impacted by early health conditions in similar ways. This finding raises the question whether investments in children remediate the effect of early deficits in health.

Schools are an obvious place to look for investments in human capital. In this section we ask whether the effect of birth weight on test scores is different for students who attend high quality versus low quality schools. Students who attend higher quality schools have higher test scores. But does a lower birth weight twin perform better relative to his counterpart if he or she attends a high quality school instead of a low quality school? In other words, does school quality remediate the effect of neonatal health deficits?

To answer this question, we measure school quality in three different ways. First, we take advantage of the fact that since 1999 the state of Florida has given each of its public schools a letter grade ranging from A (best) to F (worst). Initially, this grading system was based mainly on average proficiency rates on the FCAT. Beginning in 2002, grades were based on a combination of average FCAT proficiency rates and average student-level FCAT test score gains from year to year. In addition, we stratify schools based on average proficiency levels and average student gains from year to year. While these are only three of a wide range of ways in which one could evaluate school quality, they are sufficiently different[26] that similar findings across the three measures would provide strong evidence of the potential effects of school quality in ameliorating or exacerbating the relationships between birth weight and student cognitive development. In our analysis, therefore, we measure school quality using (1) the state awarded letter grade, (2) the school's average FCAT proficiency level during our sample period, and (3) the school's average year-to-year student FCAT gain score over our sample period.

The results of the school quality analyses are presented in table 5. The first panel of the table shows twin-pair fixed effects estimates separately for twins who attended schools that received an A, a B, and a C or below.[27] As can be seen, almost half of the sample attended schools that received a grade of A, while 28.8 percent attended a school that received a B and 22.6 percent attended a school that received a C, D or F. For reasons due either to school quality or to selection, test scores are much higher in A-rated schools than in lower-rated schools, and we also observe that twins who attend higher-rated schools tend to be born larger than those

[25] In results we report in appendix table A2, we also estimated a model with a quadratic in birth weight, which yields a positive coefficient on the linear term and a negative coefficient on the quadratic term, also suggesting a concave relationship.

[26] If we code the school grades on the scale from 0 (F) to 4 (A), we observe that state-awarded grades correlate with average school achievement at 0.68 and with growth in achievement at 0.20, while the average achievement correlates with achievement growth at 0.03.

[27] We combine C, D, and F-graded schools in this analysis because highly educated and older families, who are more likely to have twins, are more likely to live in "better" school zones than the general population, and because the state of Florida has awarded relatively few grades of D and F. In the overall population, 5.8% and 0.9% students attend D and F schools respectively, while among twins these rates are 3.4% and 0.6% respectively.

attending lower-rated schools. But while there are relationships between school grade, birth weights, and test scores, there is no monotonic relationship in the relationship between birth weight and test scores: The estimated effect of birth weight is largest among twins who attend schools receiving a grade of B (0.497). The smallest estimated effect is for twins attending A schools (0.407), and the estimate in the middle is for twins attending C/D/F schools (0.455). These coefficients are not statistically distinguishable from one another.

The second panel in the table presents results where school quality is measured based on the school's average FCAT scores. About 60 percent of the sample attended schools with scores that are above the state median average score – unsurprising given that families of twins are disproportionately older, more educated, and live in neighborhoods with higher median income. Though average test scores are certainly different in high- and low-average-test-score schools, the estimated effect of birth weight does not vary significantly. We estimate that the marginal effect of log birth weight for twins attending schools with above-average FCAT scores is 0.425. For twins attending schools with below-average FCAT scores, we estimate the effect to be 0.436.

Our estimates of the effect of log birth weight on test scores also does not vary between schools with above and below average FCAT gains. These estimates are shown in the third panel of table 5. We estimate that the marginal effect of log birth weight on test scores for twins attending a school that had below-average year-to-year gains in FCAT scores is 0.449. For twins attending a school that had above-average FCAT gains, we estimate the marginal effect of log birth weight to be 0.433.

In summary, the evidence appears to indicate that the effect of birth weight on test scores does not vary with measures of the quality of schools that a child attends. One view of this result could be that the effects of *in utero* health conditions create a ceiling to learning that cannot be remediated after the fact, at least by the time that children are of schooling age. Students spend a great deal of time in schools, and schooling is the primary formal way that human capital investment takes place during childhood. The amount (Card, 1999) and quality (Card & Krueger 1992, 1996, Krueger & Whitmore, 2001, Chetty et al., 2011a, Chetty et al., 2011b) of schooling have been shown to have significant positive impacts on earnings and other outcomes. If attending a better school does not completely remediate the effects of early health deficits on cognitive development, maybe schools currently lack the resources to fully remediate them.

An alternative view of the results is that they say that school quality does not differentially affect remediation, but leaves open the possibility that remediation *could* happen. This view is supported by a few observations. The difference in birth weights (or cognitive capacities) between twins is probably far more noticeable to parents than to classroom teachers. To a parent the outward signs of a 15 percent difference in birth weight can seem large, but to a teacher they are small relative to the variation she observes in the classroom. Even twins with large discordance in birth weight and with the resulting differences in cognitive achievement probably appear to the teacher to be quite similar to each other. Recall that the difference in achievement between the average high and low birth weight twin is far less than the difference in achievement between children born to college educated and high school dropout mothers. Given this discrepancy, it is likely that teachers treat twins very similarly. The lack of remediation may not indicate that it is impossible to remediate. Rather, it may indicate that it is not done, at least not systematically. We therefore turn to the question of whether remediation seems to be occurring in the years prior to third grade.

VIII. Birth weight gaps at kindergarten entry

The question of whether remediation for early health deficits is possible leads us to investigate impacts at an earlier age, when children spend a larger share of time with their parents. Alongside schooling, parents are the other main source of investment in children's human capital development. At ages 6-8, as children enter full time schooling, they spend on

average 30 percent less time being actively cared for by their parents than they did when they were 3-5 and 43 percent less time than when they were 0-2 (Folbre et al., 2005). The shift in time spent with parents to time spent with other adults, such as teachers, and peers (Sacerdote, 2001) suggests it may be important to gauge whether the effect of neonatal health on cognitive development is different in these early ages than during school ages. One potential reason why school quality does not appear to affect the relationship between neonatal health and cognitive development may be that differences in factors correlated with poor neonatal health might be minimally perceptible to teachers and school administrators who interact with a wide range of children, but more noticeable to parents who make cross-child observations. And beyond learning about the roles of adult investment and remediation, there is an inherent value in documenting the developmental effects of early health on cognitive development. To provide further evidence on these matters, we extend the age at which we measure outcomes back to when children enter kindergarten, which typically happens at age 5 or 6.

In various years between 1998 and 2008, Florida performed universal kindergarten readiness screening and recorded this screening in its Education Data Warehouse. From 1998 through 2001 all kindergarten entrants were screened with the School Readiness Checklist (SRC), a list of 17 expectations for kindergarten readiness. Subsequently, kindergarten entrants were screened with the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), and beginning in 2006 the results of this screening were collected and recorded by the Florida Department of Education.[28] DIBELS rates children's letter sound recognition and letter naming skills and categorizes children as above average, low risk, moderate risk or high risk. In our data, 82.1 percent of children were deemed ready according to the earlier SRC screen, and a very similar 83.8 percent of children were deemed either above average or low risk according to the DIBELS.

These two kindergarten readiness screens are clearly different outcome measures than the FCAT. However, they are highly predictive of later test scores. The average third through eighth grade FCAT score for children deemed ready to start school ranges from 0.147 to 0.280 depending on the cohort, while the average FCAT score for those not ready to start school ranges from -0.655 to -0.501, depending on the cohort.[29] The correlation with subsequent outcomes is present within families as well: In twin fixed effects models, we estimate that average third through eighth grade FCAT score differences between twins who were ready versus not ready for kindergarten range from 0.22 to 0.32 of a standard deviation, depending on the cohort.

Turning to estimates of the effect of birth weight on kindergarten readiness, we first present results from models like the baseline specification reported above, but replacing FCAT scores with a dummy variable for being deemed ready for kindergarten. These estimates show whether the effect of birth weight on cognitive skills is present at age 5. We next analyze how the magnitude of these effects compare with the magnitudes we find for third through eighth grade test scores.

We begin by simply presenting the estimated effects of log birth weight on the different variants of the kindergarten readiness assessment. The first three rows of table 6 present the coefficients on log birth weight in three different cases – the initial school readiness checklist cohorts, the DIBELS cohorts, and the two kindergarten readiness assessments pooled together. One observes that a 10 percent increase in birth weight is associated with a 0.67 percentage point increase in being deemed ready for kindergarten according to the school readiness checklist, and

[28] For more details about the structure and interpretation of DIBELS, see for instance Hoffman et al. (2009).

[29] While the kindergarten readiness assessments, and especially the DIBELS, focuses more on pre-literacy skills than on numeracy skills, these kindergarten readiness assessments are predictive of both later reading and mathematics achievement. The gap between kindergarten-ready and kindergarten-unready children ranges between 0.609 and 0.823, depending on cohort, for reading FCAT scores and between 0.680 and 0.868, depending on cohort, for mathematics FCAT scores.

a 10 percent increase in birth weight is associated with a 1.15 percentage point increase in kindergarten readiness according to the DIBELS. When we pool the two sets of cohorts, these figures average to a 0.86 percentage point increase. All of these estimates are statistically distinct from zero at conventional levels.

These results make it clear that the effect of neonatal health on cognitive development that we documented for ages 9-14 is present by age 5. How does the magnitude of the effects at kindergarten entry compare with the magnitude of the effects in grades three through eight? The kindergarten readiness outcome is a binary indicator and the grade three through eight FCAT scores are continuous standardized *z*-scores. To compare the two outcomes, we transformed the FCAT scores into a binary indicator. To match the kindergarten readiness screens, we created a dummy variable that equals one if the student's FCAT score is above the 17th percentile (the threshold that corresponds to the fraction of kindergarteners deemed not ready in the pooled SRC and DIBELS samples). We then estimated the baseline log birth weight fixed effects model for the pooled third through eighth grade sample and then separately by grade. These threshold-based estimates are presented in the fourth through sixth rows of table 6. In order to ensure that we are comparing kindergarten readiness to test scores for the same children, we limit our comparisons to children for whom we observe a balanced panel of kindergarten readiness assessments and FCAT scores between third and eighth grade.

The three rows differ in terms of how late into school a child must be observed to be included in the analysis. The fourth row includes all children observed at both kindergarten and third through eighth grade, so includes only students who took the SRC and none who took the DIBELS, while the sixth row includes children observed at both kindergarten and third grade, so students who took the DIBELS and SRC are more evenly represented.[30] As can be seen, the coefficient estimates on log birth weight in models where the kindergarten readiness assessment is the dependent variable (ranging from 0.057 to 0.093) are considerably smaller than the coefficient estimates in models in which the discretized FCAT score is the dependent variable (ranging from 0.159 to 0.181 if limited to third grade, and from 0.146 to 0.167 if all grades are pooled.) These differences are statistically significant at conventional levels, which could be interpreted as an increase in the effect of birth weight on test scores between kindergarten and third grade.

There is reason to believe, however, that the effects of birth weight on outcomes remain roughly constant between kindergarten and third grade (and therefore, through at least eighth grade.) Recall that the SRC and DIBELS reflect somewhat different skills; while the SRC reflects numeracy, literacy, and behavioral skills, the DIBELS is explicitly a pre-literacy assessment. The fact that the estimated effects of birth weight on DIBELS are closer in magnitude to the estimated effects on FCAT scores than are the estimated effects of birth weight on SRC could mean that birth weight has a greater effect on cognitive readiness in kindergarten than it does on social and emotional readiness in kindergarten. Indeed, given that the estimated effects of birth weight on discretized reading and mathematics FCAT scores are the same whether we consider the set of students who took the SRC or the set of students who took the DIBELS (see rows 7 and 8 in table 6),[31] and given that the estimated birth weight effects on reading, math, and DIBELS scores are all reasonably in line with one another in the panel that considers them all simultaneously (see

[30] None of the cohorts for whom we have DIBELS scores are yet old enough to have reached 8th grade. The oldest of the DIBELS cohorts was in 6th grade in 2011-12, the most recent year of our testing data.

[31] The coefficient estimates in the third column of table 6 are not always between the coefficient estimates in the fifth and sixth columns because we discretized reading, math, and average test scores separately for each column. If we use as our dependent variable in the third column the average of the discretized reading and discretized math scores, rather than the discretized average reading and math score, then the estimated effect of log birth weight is exactly midway between the estimated effects on reading and math. The discrepancy occurs because reading and math scores are not perfectly correlated.

row 8 in table 6), the evidence suggests that the effects of birth weight on *cognitive* skills remain steady from kindergarten through schooling. Indeed, there exists some evidence that the cognitive differences associated with birth weight seen throughout schooling are also present at about the same magnitude in early childhood: In a study comparing a much smaller set of twins born in 2001 in the ECLS-B, Hart (2008) finds estimated effects of birth weight on the Bayley Mental Scale that are remarkably similar in effect size to those presented in our paper.

Taken together, our findings indicate that the effects of poor neonatal health on cognitive development appear to be largely fixed by the time children enter kindergarten, and that the pattern of results are consistent with the notion that parental inputs and neonatal health are complements rather than substitutes. We should also point out that there is evidence that parents actively make different decisions regarding their twins' early childhood experiences, suggesting that parents recognize developmental differences in their children and seek to remediate these differences in early childhood. In our data, it is reasonably common for parents to send one twin to preschool but not the other (true in 7.6 percent of twin pairs and 8.8 percent of twin pairs in which the birth weight discordance is greater than 20 percent). In 9.3 percent of twin pairs (10.5 percent of twin pairs with discordance greater than 20 percent) parents choose different preschool arrangements for their twins – either sending one twin to preschool but not the other, or sending both twins to preschool but only one to privately-financed preschool. And in just under one percent of cases (1.3 percent of twin pairs with discordance greater than 20 percent) parents "redshirt" one twin but not the other – sending the twins to kindergarten at different times.[32]

Coupled with the evidence on attempted parental remediation of disadvantage in the United States (Hsin, 2012) and the evidence on differential parental time use in early childhood versus early elementary grades (Folbre et al., 2005), the evidence is consistent with the notion that parental remediation patterns could help to lower the negative effects of poor neonatal health in early childhood. But the finding of a significant effect of birth weight on kindergarten readiness in twin-comparison models indicates that there is an apparent limit to the degree of remediation that is likely, and that a portion of the biological factors are apparently difficult to overcome.

IX. Conclusion

Using a unique population-level data source from Florida, we present the first look at the effects of poor neonatal health on child cognitive development in a western developed context, provide the first study of the differential effects on different demographic and socio-economic groups, and offer the first exploration of the degree to which school quality might influence these effects. Our results are remarkably consistent: Twins with higher birth weights enter school with a cognitive advantage that appears to remain stable through the elementary and middle school years. The estimated effects of low birth weight are present for children of highly-educated and poorly-educated parents alike, for children of both young and old mothers, and for children of all races and ethnicities, parental immigration status, parental marital status, and the like. The estimated effects are just as pronounced for students attending highly-performing public schools (measured in a variety of ways) as they are for students attending poorly-performing public schools. These results strongly point to the notion that the effects of poor neonatal health on adult outcomes are largely determined early – in early childhood and the first years of elementary school.

There exists circumstantial evidence to suggest that these biological impediments may be remediated, but neither schools nor parents are able to fully remediate these factors. It is the case that children with poor neonatal health who come from highly-educated families perform much

[32] In cases of differential redshirting, parents are slightly more likely to redshirt the lighter twin than they are to redshirt the heavier twin.

better than those with good neonatal health who come from poorly-educated families, indicating that "nurture" can at least partially overcome "nature." While what exactly parents do to successfully remediate biological disadvantage and what schools and parents could potentially do in early childhood and the early elementary grades and beyond to continue to remediate these issues are open questions, this study provides numerous indications that poor neonatal health establishes a stable trajectory for children's cognitive development.

Bibliograpy

Aarnoudse-Moens, Cornelia, Nynke Weisglas-Kuperus, Johannes van Goudoever, and Jaap Ooseterlann. 2009. "Meta-Analysis of Neurobehavioral Outcomes in Very Preterm an/or Very Low Birth Weight Children", Pediatrics 124(2): 717-728.

Almond, Douglas, Kenneth Y. Chay and David S. Lee. 2005. "The Costs of Low Birth Weight", Quarterly Journal of Economics 120(3): 1031-1083

Almond, Douglas, Lena Edlund and Marten Palme. 2009. "Chernobyl's Subclinical Legacy: Prenatal Exposure to Radioactive Fallout and School Outcomes in Sweden", Quarterly Journal of Economics 124(4): 1729-1772

Almond, Douglas, and Janet Currie. 2011. "Killing Me Softly: The Fetal Origins Hypothesis", Journal of Economic Perspectives 25(3): 153-172

Ananth, D.C. and S.P. Chauhan. 2012. "Epidemiology of Twinning in Developed Countries", Seminars in Perinatology 36: 156-161.

Behrman, Jere, and Mark R. Rosenzweig. 2004. "Returns to Birthweight", Review of Economics and Statistics 86(2): 586-601

Bharadwaj, Prashant, Juan Eberhard, and Christopher Neilson. 2010. "Do Initial Endowments Matter Only Initially? Birth Weight, Parental Investments and Academic Achievement in School", working paper, University of California-San Diego

Bitler, Marianne. 2008. "Effects of Increased Access to Infertility Treatment on Infant and Child Health: Evidence from Health Insurance Mandates", working paper, University of California-Irvine

Black, Sandra E., Paul J. Devereux and Kjell G. Salvanes. 2007. "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes", Quarterly Journal of Economics 122(1): 409-439

Camacho, Adriana. 2008. "Stress and Birth Weight: Evidence from Terrorist Attacks", American Economic Review: Papers and Proceedings, 98(2): 511-515

Card, David. 1999. "The Causal Effect of Education on Earnings", in: Orley Ashenfelter and David Card (eds.), Handbook of Labor Economics 3A, Amsterdam: Elsevier

Card, David, and Alan B. Krueger. 1996. "Labor Market Effects of School Quality: Theory and Evidence", in: Gary Burtless (eds.), The Link Between Schools, Student Achievement, and Adult Success, Washington D.C.: Brookings Institution

Card, David, and Alan B. Krueger. 1992. "School Quality and Black-White Relative Earnings: A Direct Assessment", Quarterly Journal of Economics 107(1): 151-200

Case, Anne, Angela Fertig and Christina Paxson. 2005. "The Lasting Impact of Childhood Health and Circumstance", Journal of Health Economics, 24: 365-389

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach and Danny Yagan. 2011a. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star", Quarterly Journal of Economics 126(4): 1593-1660

Chetty, Raj, John N. Friedman and Jonah Rockoff. 2011b. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood", NBER Working Paper # 17699

Cheung, V.Y., A.D. Bocking, and O.P. Dasilva. 1995. "Preterm Discordant Twins: What Birth Weight Difference is Significant?" American Journal of Obstetrics and Gynecology 172(3): 955-959.

Folbre, Nancy, Jayoung Yoon, Kade Finnoff and Allison Sidle Fuligni. 2005. "By What Measure? Family Time Devoted to Children in the United States," *Demography* 42(2): 373-390

Hart, Cassandra. 2008. "Parenting and Child Cognitive and Socioemotional Development: A Longitudinal Twin Differences Study", working paper, Northwestern University

Hoffman, Amy R., Jeanne E. Jenkins and Kay S. Dunlap. 2009. "Using DIBELS: A Survey of Purposes and Practices", Reading Psychology 30(1): 1-16

Hsin, Amy. 2012. "Is Biology Destiny? Birth Weight and Differential Parental Treatment", Demography 49(4): 1385-1405

Krueger, Alan B. and Diane Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR", Economic Journal 111(468): 1-28

Lam, Lucia L., Eldon Emberly, Hunter B. Fraser, Sarah M. Neumann, Edith Chen, Gregory E. Miller, and Michael S. Kobor. 2012. "Factors Underlying Variable DNA Methylation in a Human Community Cohort", Proceedings of the National Academy of Sciences 109(Supplement 2): 17253-17260.

Luu, T.M. and B. Vohr. 2009. "Twinning on the Brain: The Effect on Neurodevelopmental Outcomes", American Journal of Medical Genetics, Part C: Seminars in Medical Genetics 151C(2): 142-147

Miller, Gregory E., Edith Chen, Alexandra K. Fok, Hope Walker, Alvin Lim, Erin F. Nicholls, Steve Cole, and Michael S. Kobor. 2009. "Low Early-Life Social Class Leaves a Biological Residue Manifested by Decreased Glucocorticoid and Increased Proinflammatory Signaling", Proceedings of the National Academy of Sciences 106(34): 14716-14721

Neal, Derek A. and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences", Journal of Political Economy 104(5): 869-895

Oreopoulos, Philip, Mark Stabile, Randy Walld and Leslie L. Roos. 2008. "Short-, Medium-, and Long-Term Consequences of Poor Infant Health. An Analysis Using Siblings and Twins", Journal of Human Resources 43(1): 88-138

Rosenzweig, Mark R., and Junsen Zhang. 2009. "Do Population Control Policies Induce More Human Capital Investment? Twins, Birth Weight and China's "One-Child" Policy", Review of Economic Studies 76: 1149-1174

Rosenzweig, Mark R., and Junsen Zhang. 2012. "Economic Growth, Comparative Advantage, and Gender Differences in Schooling Outcomes: Evidence from the Birthweight Differences of Chinese Twins", Yale University Economics Department Working Paper #98

Royer, Heather. 2009. "Separated at Girth: US Twin Estimates of the Effects of Birth Weight", American Economic Journal: Applied Economics 1(1): 49-85

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates", Quarterly Journal of Economics 116(2): 681-704

Torche, Florencia, and Ghislaine Echevarria. 2011. "The Effect of Birthweight on Childhood Cognitive Development in a Middle-Income Country", International Journal of Epidemiology 40(4): 1008-1018

Vergani, P., A. Locatelli, M. Ratti, A. Scian, E. Pozzi, J.C. Pezzullo, and A. Ghidini. 2004. "Preterm Twins: What Threshold of Birth Weight Discordance Heralds Major Adverse Neonatal Outcome?" American Journal of Obstetrics and Gynecology 191(4): 1441-1445.

Zwicker, J.G. and S.R. Harris. 2008. "Quality of Life of Formerly Preterm and Very Low Birth Weight Infants from Preschool Age to Adulthood: A Systematic Review", Pediatrics 121(2): 366-376.

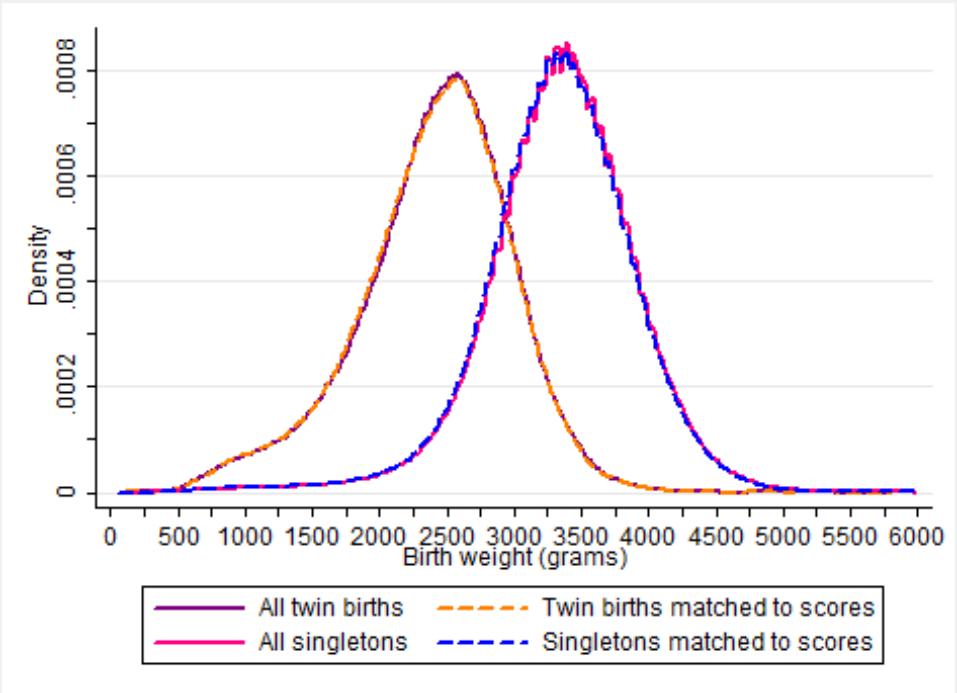
FIGURES

Figure 1. Discordance in birth weight between twins born in Florida between 1992 and 2002



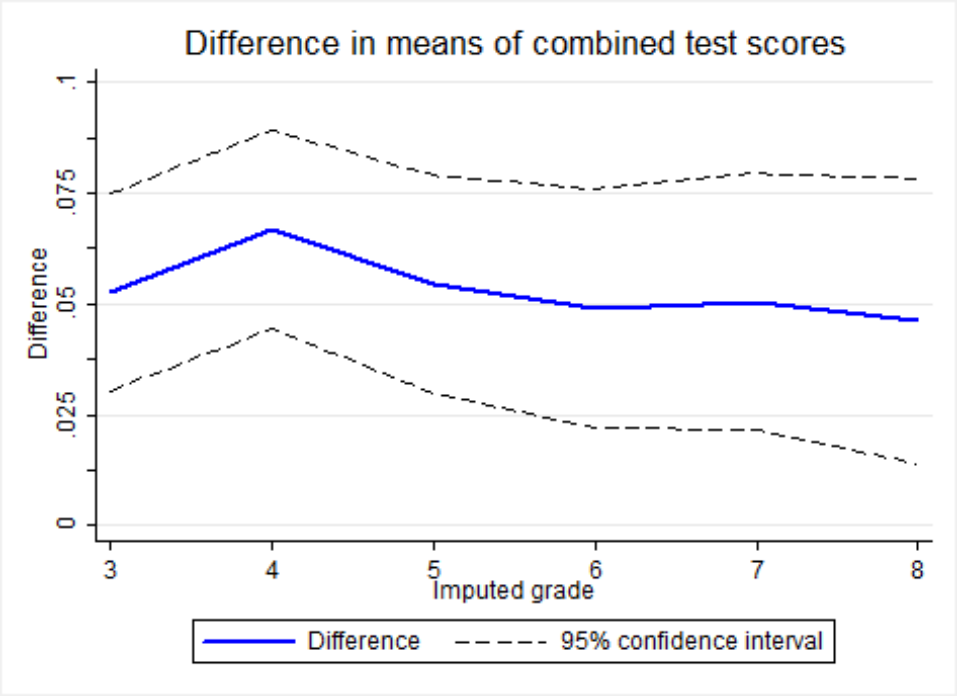
Note: Figure 1 plots kernel density distributions of within-twin-pair difference in birth weight for all twin births in Florida (solid pink line) between 1992 and 2002 and twin births who were born in Florida and were successfully matched to Florida public school records (dashed blue line). Distributions are censored at 2000 grams for the sake of clarity, which removes 6 and 3 twin pairs respectively.

Figure 2. Difference in birth weight distributions between singletons and twins born in Florida between 1992 and 2002



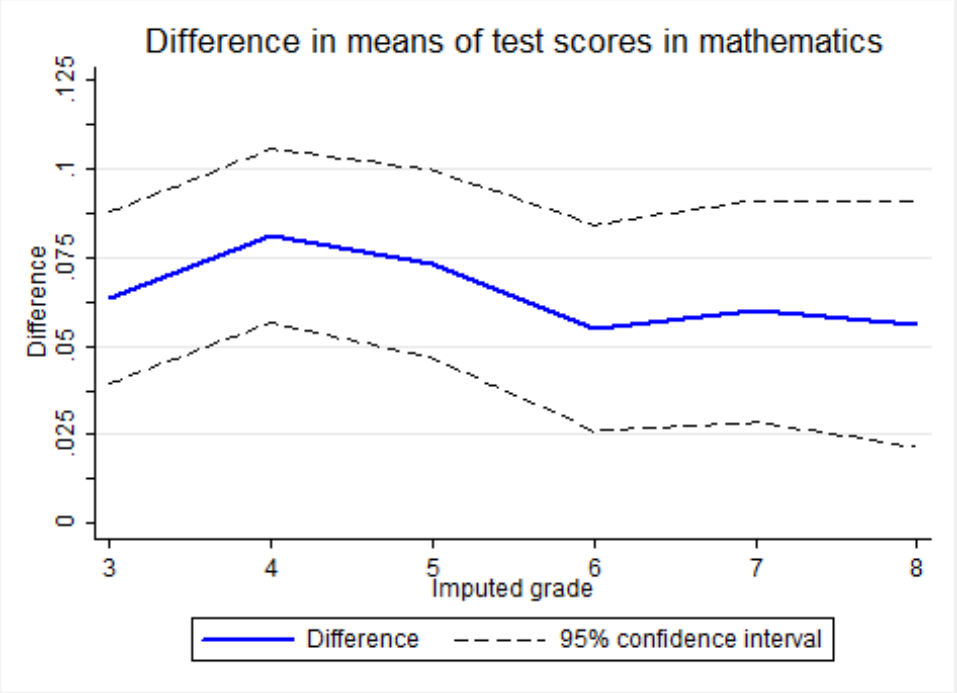
Note: Figure 2 plots kernel density distributions of infant birth weight for all singletons (solid pink line) and twins (solid purple line) born in Florida between 1992 and 2002 as well as infant birth weight distribution of singletons (dashed blue line) and twins (dashed orange line) that were successfully matched to Florida public school records.

Figure 3. Average within-twin-pair difference in test scores between heavier and lighter twins



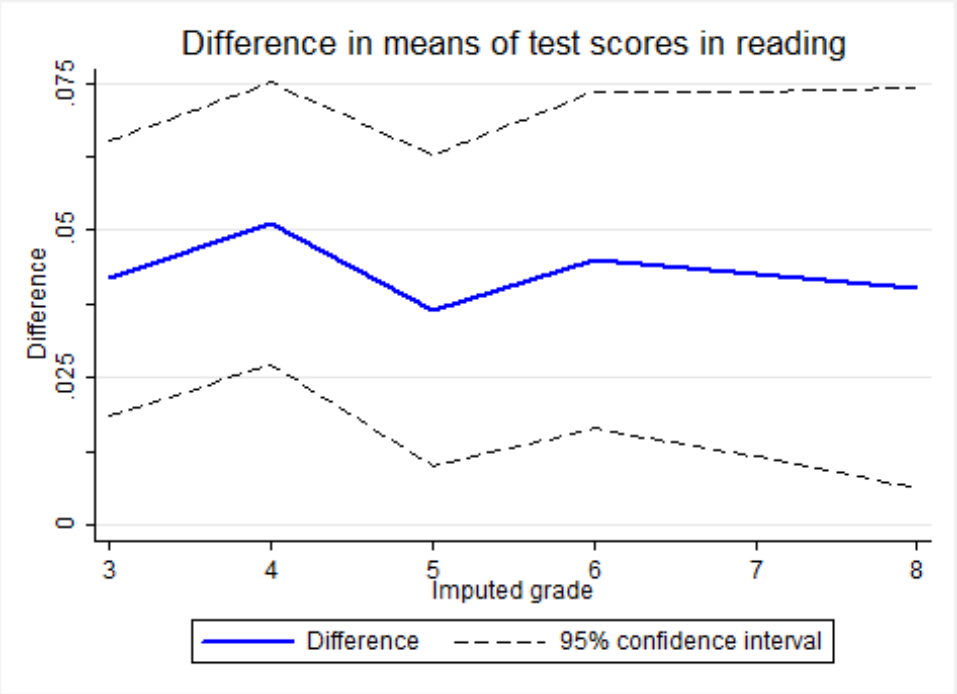
Note: Figure 3 plots difference between the mean test score of heavier and lighter twin from each pair in each grade and the respective 95% confidence interval of this difference. Mean test score is constructed as an average of scores in mathematics and reading for each individual in each grade where we observe both twins. If score in mathematics is not available then only reading is used and vice versa. In each grade we create an average of scores for heavier and lighter twins and then calculate the difference between the two.

Figure 4. Average within-twin-pair difference in mathematics between heavier and lighter twins



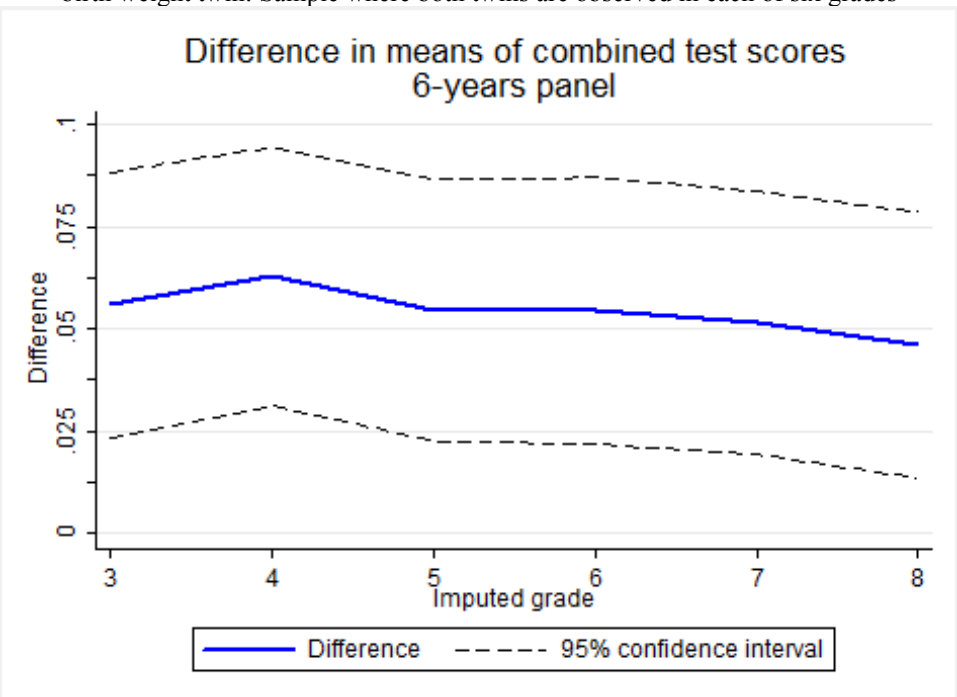
Note: Figure 4 plots difference between the test score in mathematics of heavier and lighter twin from each pair in each grade and the respective 95% confidence interval of this difference. In each grade we create an average of scores for heavier and lighter twins and then calculate the difference between the two.

Figure 5. Average within-twin-pair difference in reading between heavier and lighter twins



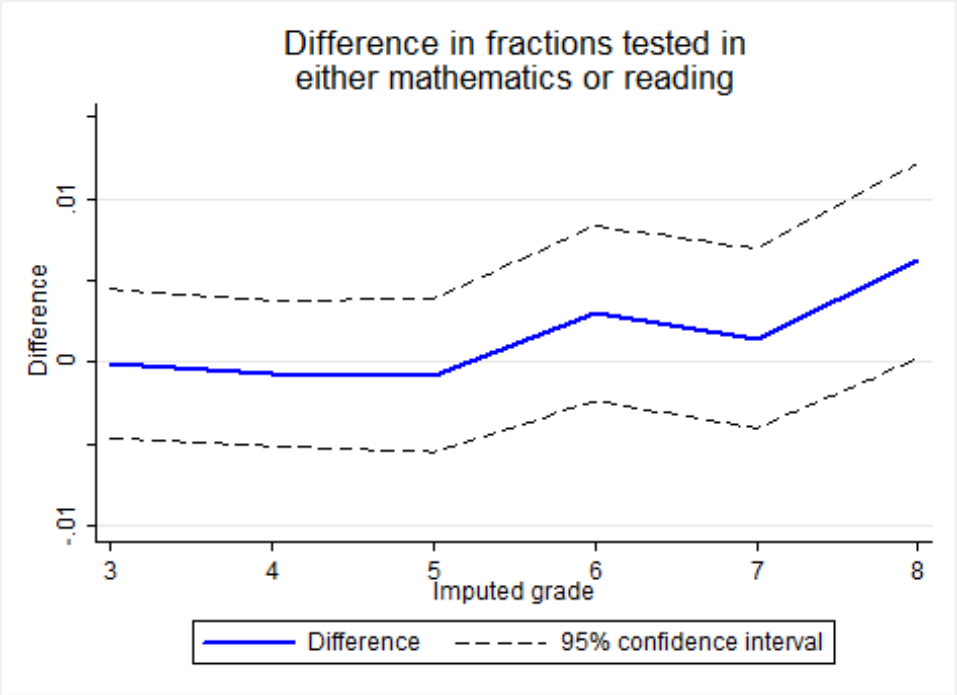
Note: Figure 5 plots difference between the test score in reading of heavier and lighter twin from each pair in each grade and the respective 95% confidence interval of this difference. In each grade we create an average of scores for heavier and lighter twins and then calculate the difference between the two.

Figure 6. Average within twin pair difference in test scores between the higher birth weight and the lower birth weight twin: Sample where both twins are observed in each of six grades



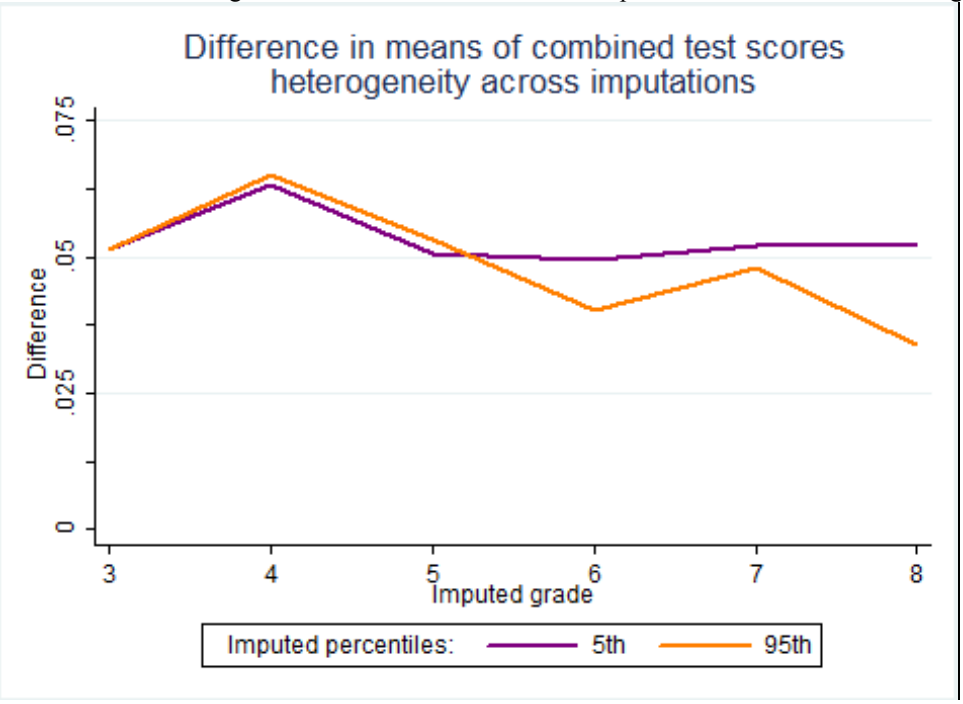
Note: Figure 6 plots the same difference as Figure 3 but for a 6-year panel of twin-pairs i.e., we restrict the sample only to individuals where we observe both twins mean test scores from grade 3 to grade 6.

Figure 7. Difference in fraction of lighter and heavier twins tested in either mathematics or reading

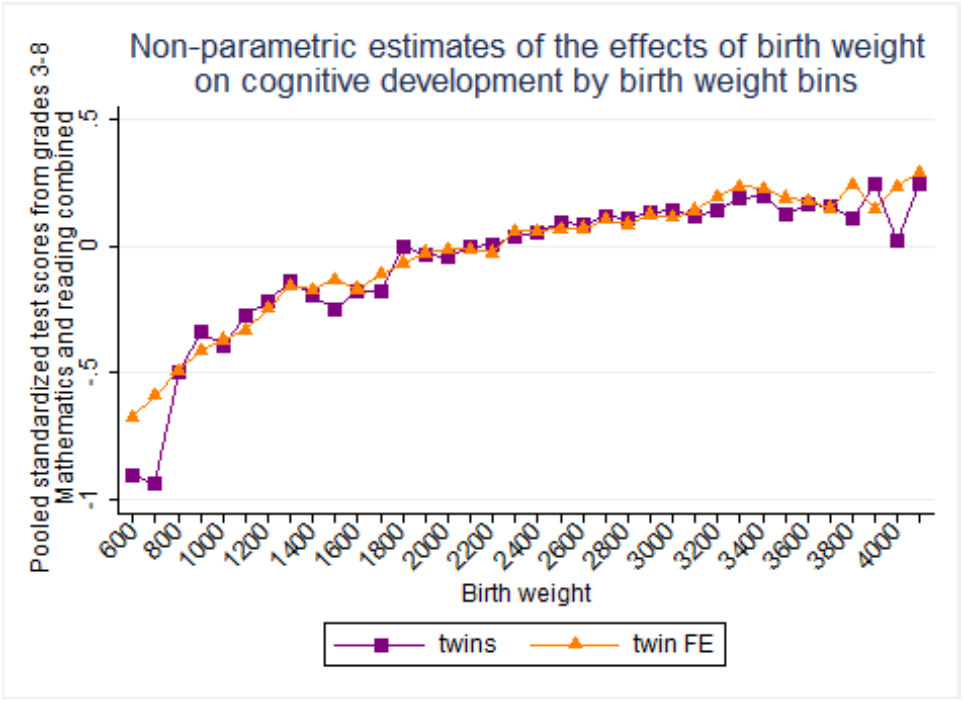


Note: Figure 7 plots difference and its 95% confidence interval of fraction of heavier and lighter twins attending each grade. We start with all twin pairs old for grade where at least one twin has been successfully matched to Florida public schools in 3rd grade. For each grade we then calculate the fraction of heavier and lighter individuals attending given grade and take the difference between the two.

Figure 8. Differences across grades with 5th and 95th test score imputations for twins with missing scores

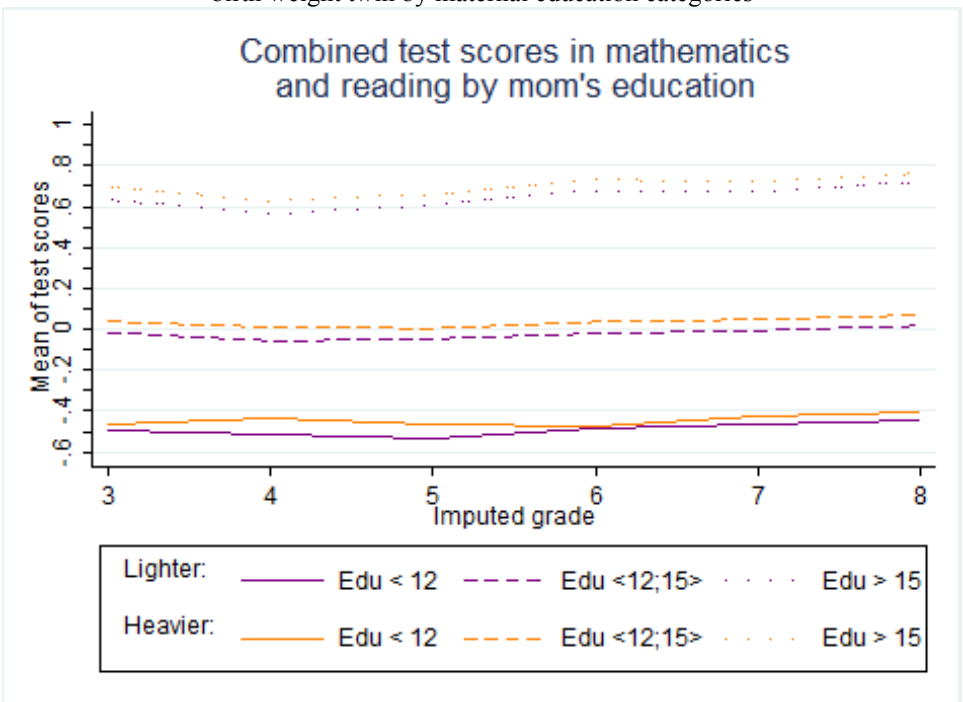


Note: Figure 8 plots two sets of differences calculated in the same way as in figure 3 but where we substitute the missing individual scores within twin pairs with either the 5th (solid purple line) or 95th (solid orange line) percentile of test scores in that grade.



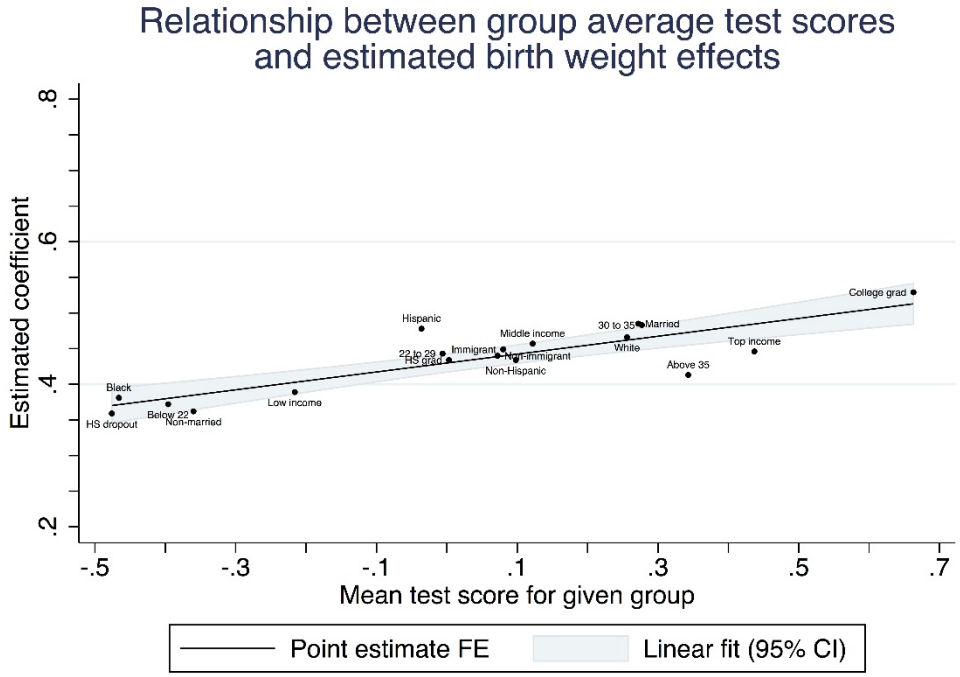
Note: Figure 9 plots coefficients from OLS (purple solid line) and twin-FE (orange solid line) models where the dependent variable is the mean of pooled grades three through eight combined mathematics and reading test scores for each individual and the independent variables are indicators for 37 weight bins corresponding to each individual birth weight. No additional controls are included in the models.

Figure 10. Average within twin pair difference in test scores between the higher birth weight and the lower birth weight twin by maternal education categories



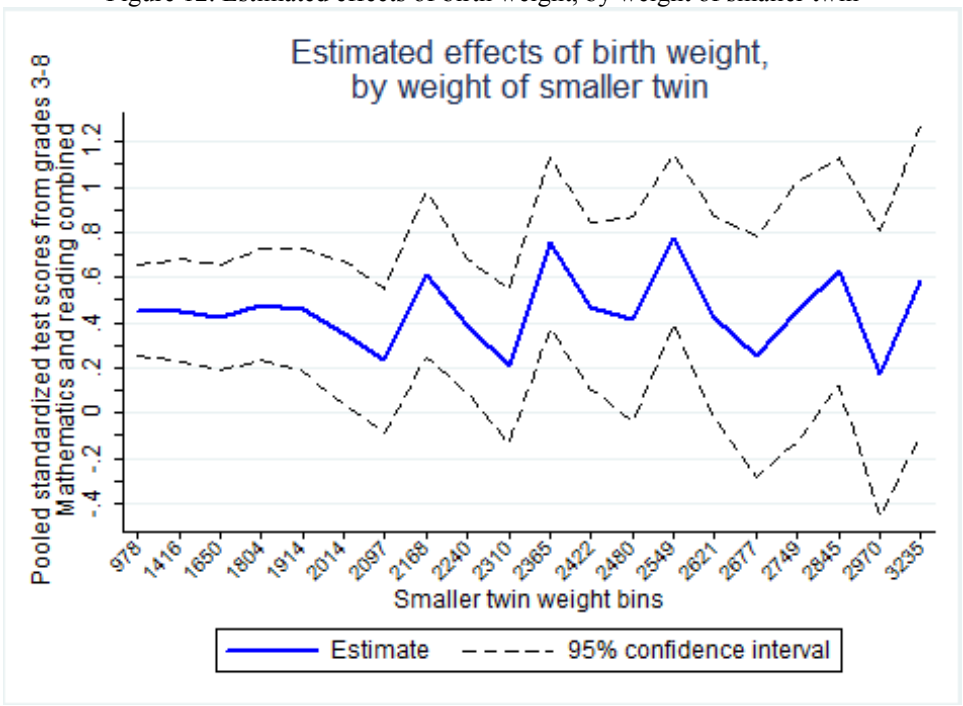
Note: Figure 10 plots means of combined mathematics and reading test scores for lighter and heavier twins from each pair stratified by maternal education. Purple lines correspond to averages for lighter while orange lines correspond to heavier twins. Solid lines present means for high school drop-out mothers, dashed lines present means for children of mothers with high school diploma or some college while dotted lines present means for college graduates.

Figure 11. Average test scores among groups and estimated birth weight effects



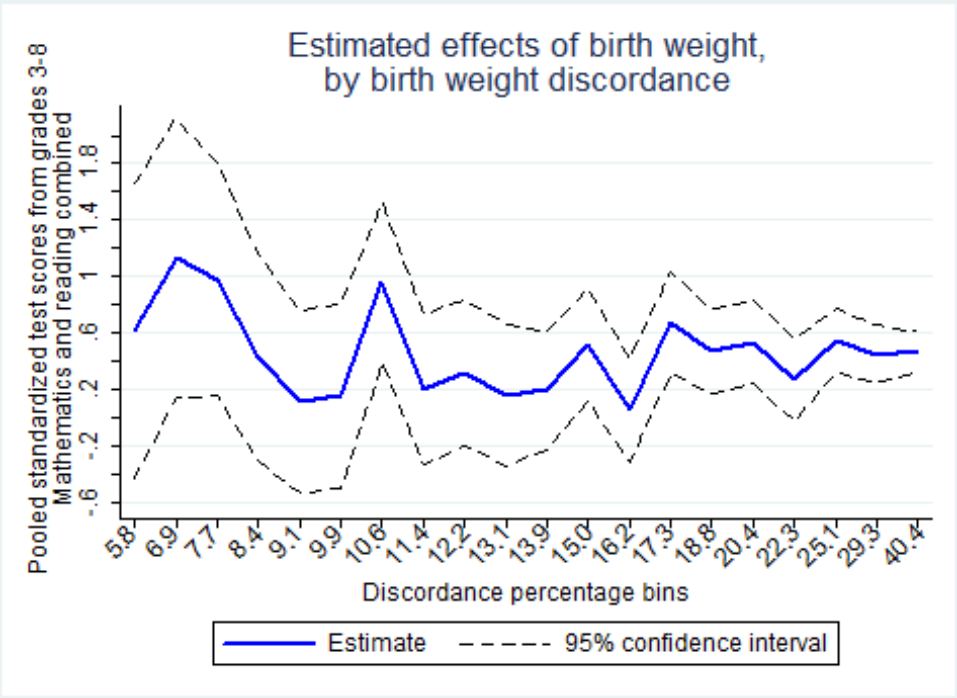
Note: Figure 11 plots the pooled coefficients presented in rows 4 to 10 in table 3 against the mean tests scores for each of the groups presented in column (2) of this table. We also fit the linear prediction with 95% confidence interval around it. Labels correspond to each of the studied groups.

Figure 12. Estimated effects of birth weight, by weight of smaller twin



Note: Figure 12 plots coefficients and 95% confidence intervals from a twin FE regression where the dependent variable is the mean of pooled grades three to eight combined mathematics and reading test scores for each individual and the independent variables are 20 interactions corresponding to the product of log birth weight with indicators for 20 bins reflecting lighter twin percentiled birth weight. The regression additionally controls for infant gender and birth order within-twin pair. Heteroskedasticity robust standard errors are used to calculate the 95% confidence interval. Numbers on the x-axis correspond to the mean birth weight discordance in each of the 20 bins.

Figure 13. Estimated effects of birth weight, by birth weight discordance



Note: Figure 13 plots coefficients and 95% confidence intervals from a twin FE regression where the dependent variable is the mean of pooled grades three to eight combined mathematics and reading test scores for each individual and the independent variables are 20 interactions corresponding to the product of log birth weight with indicators for 20 bins reflecting discordance in birth weight between twins. The regression additionally controls for infant gender and birth order within-twin pair. Heteroskedasticity robust standard errors are used to calculate the 95% confidence interval. Numbers on the x-axis correspond to the mean birth weight in each bin of lighter twin birth weight.

TABLES

Table 1. Representativeness of the Florida twin population

Maternal attribute	(1) Full population of births	(2) Population of kids matched to Florida school records	(3) Population of kids with a third-grade test score	(4) Population of twins with a third grade test score
Black	22.6	24.8	25.7	25.9
Hispanic	23.0	23.3	23.8	18.0
High school dropout	20.9	22.5	23.4	15.5
High school graduate	58.6	60.0	60.8	61.5
College graduate	20.5	17.5	15.8	23.0
Age 21 or below	22.0	23.6	24.2	14.4
Age between 22 and 29	42.2	42.2	42.2	40.2
Age between 30 and 35	26.0	24.8	24.5	31.8
Age 36 or above	9.8	9.4	9.1	13.6
Foreign-born	23.5	22.9	23.1	18.0
Married at time of birth	64.8	62.2	61.0	68.3
Number of children	2,047,663	1,652,333	1,326,004	28,564

Note: The first column presents fractions in total population of children born in Florida between 1992 and 2002. The second column presents fractions in total population of children born between 1992 and 2002 linked to Florida school records. The third column presents fractions in total population of children born between 1992 and 2002 for whom we observe a third grade test score. Fourth column presents fractions in total population of twin pairs born between 1992 and 2002 for whom we observe third grade test scores.

Table 2. Estimated effects of birth weight on cognitive development

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pooled		Imputed grade: Twin FE models					
	OLS	Twin FE	3	4	5	6	7	8
	Average of mathematics and reading:							
Ln(birth weight)	0.310***	0.441***	0.442***	0.526***	0.430***	0.426***	0.386***	0.373***
	(0.019)	(0.029)	(0.043)	(0.045)	(0.047)	(0.053)	(0.056)	(0.061)
R^2	0.180	0.751	0.825	0.813	0.825	0.831	0.826	0.834
N	127,156		28,564	26,628	23,056	19,408	16,246	13,254
	Mathematics:							
Ln(birth weight)	0.387***	0.497***	0.473***	0.579***	0.533***	0.490***	0.410***	0.429***
	(0.020)	(0.032)	(0.051)	(0.051)	(0.052)	(0.061)	(0.067)	(0.070)
R^2	0.159	0.708	0.806	0.793	0.809	0.807	0.801	0.811
N	126,542		28,496	26,552	22,986	19,332	16,136	13,040
	Reading:							
Ln(birth weight)	0.230***	0.392***	0.415***	0.467***	0.328***	0.372***	0.370***	0.349***
	(0.019)	(0.031)	(0.048)	(0.051)	(0.055)	(0.059)	(0.062)	(0.069)
R^2	0.158	0.697	0.795	0.787	0.793	0.805	0.798	0.806
N	126,706		28,470	26,520	22,976	19,348	16,210	13,182

Note: Columns (1) and (2) present pooled grade three through eight results for OLS and twin-FE models. Columns (3) to (8) present twin-FE estimates separately for each of the 6 grades. Each coefficient comes from a separate regression. Sample sizes reflect number of individual observations in each regression and only twin pairs where both twins are observed with test scores in each grade are included. The dependent variable is an average test scores in mathematics and reading. If the test score in mathematics is not available then reading is included and vice versa. The main variable of interest is natural logarithm of birth weight. The remaining independent variables in twin-FE models include infant gender wand within-twin pair birth order. OLS estimates further controls for infant birth month and year, indicators for maternal age (each for one year) and education (high school dropout, high school graduate, college graduate). Standard errors in pooled regressions (columns (1) and (2)) are clustered at individual level; heteroskedasticity robust standard errors are calculated in columns (3) to (8) where there is just one observation per individual.

Table 3. Effects of birth weight on cognitive development by child and mother characteristics

Characteristic	Sample	(1) % population	(2) Mean test score	(3) Mean (SD) birth weight	(4) Pooled twin FE estimate	(5) p-value of difference
(1) Children gender composition	Same sex	68.2	0.073	2405 (568)	0.447*** (0.032)	0.773
	Opposite sex	31.8	0.076	2454 (557)	0.427*** (0.062)	
(2) Gender	Boys	49.6	0.048	2473 (571)	0.449*** (0.052)	0.941
	Girls	50.4	0.099	2369 (555)	0.444*** (0.040)	
(3) Same-sex composition	Girl-Girl	35.2	0.100	2359 (561)	0.444*** (0.039)	0.940
	Boy-Boy	33.7	0.044	2452 (572)	0.449*** (0.051)	
(4) Maternal race (N=14 357)	White	72.0	0.256	2457 (554)	0.466*** (0.034)	0.223
	Black	26.1	-0.466	2318 (585)	0.381*** (0.061)	
(5) Maternal ethnicity	Non-Hispanic	82.0	0.098	2413 (565)	0.434*** (0.033)	0.518
	Hispanic	18.0	-0.036	2454 (564)	0.478*** (0.059)	
(6) Maternal immigration history	Non-immigrant	82.0	0.072	2413 (564)	0.440*** (0.033)	0.899
	Immigrant	18.0	0.080	2451 (570)	0.449*** (0.058)	
(7) Maternal education	< 12	15.8	-0.476	2338 (570)	0.359*** (0.070)	0.163
	<12; 15>	61.4	0.003	2430 (563)	0.434*** (0.038)	
	> 15	22.8	0.663	2451 (562)	0.529*** (0.059)	
(8) Zip code median income (N=11 868)	Bottom	36.7	-0.216	2393 (567)	0.389*** (0.057)	0.657
	Middle	33.1	0.122	2409 (568)	0.457*** (0.054)	
	Top	30.2	0.437	2435 (561)	0.446*** (0.059)	
(9) Maternal marital status (N=14 583)	Non-married	31.8	-0.360	2336 (574)	0.362*** (0.057)	0.064
	Married	67.6	0.272	2458 (556)	0.485*** (0.033)	
(10) Maternal age at birth of children	<= 21	14.7	-0.396	2269 (574)	0.372*** (0.086)	0.698
	<22; 29>	40.2	-0.006	2419 (561)	0.443*** (0.044)	
	<30; 35>	31.6	0.277	2465 (557)	0.483*** (0.052)	
	>= 36	13.5	0.343	2479 (559)	0.413*** (0.078)	

Note: Descriptive statistics for each group in columns (1) to (2). Column (1) presents the fraction for each group within total population of twin pairs used in the analysis (born in Florida between 1992 and 2002 and successfully matched to Florida public schools). Columns (2) and (3) present mean combined mathematics and reading test scores and mean (SD) of birth weight for each studied group respectively. Column (4) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Column (5) presents the joint significance test for the analyzed groups in fixed effects model from column (4). Sample size: 127,156. Standard errors are clustered at the individual level.

Table 4. Sensitivity of results to model specification

Sample	(1) Pooled	(2)	(3)	(4)	(5)	(6)	(7)
		Imputed grade					
		3	4	5	6	7	8
(1)	0.441***	0.442***	0.526***	0.430***	0.426***	0.386***	0.373***
ln(birth weight)	(0.029)	(0.043)	(0.045)	(0.047)	(0.053)	(0.056)	(0.061)
(2)	0.526***	0.575***	0.656***	0.464***	0.528***	0.409***	0.427***
Both twins above 2500g	(0.073)	(0.110)	(0.115)	(0.120)	(0.126)	(0.144)	(0.152)
(3)	0.473***	0.470***	0.569***	0.495***	0.455***	0.339***	0.419***
Both twins below 2500g	(0.046)	(0.066)	(0.069)	(0.074)	(0.087)	(0.089)	(0.098)
(4)	0.517***	0.390***	0.537***	0.572***	0.591***	0.499***	0.579***
Both twins 1500g-2499g	(0.062)	(0.092)	(0.097)	(0.105)	(0.114)	(0.120)	(0.131)
(5)	0.572***	0.604***	0.714***	0.721***	0.485**	0.360	0.295
Both twins <1500g	(0.114)	(0.176)	(0.157)	(0.195)	(0.205)	(0.226)	(0.270)
(6)	0.186***	0.185***	0.223***	0.178***	0.180***	0.169***	0.155***
Birth weight in 1000g	(0.013)	(0.019)	(0.019)	(0.020)	(0.023)	(0.024)	(0.026)
(7)	0.198***	0.196***	0.234***	0.191***	0.193***	0.177***	0.171***
Birth weight	(0.013)	(0.019)	(0.020)	(0.021)	(0.023)	(0.025)	(0.027)
Birth weight * (birth weight -	-0.105***	-0.117***	-0.105***	-0.114***	-0.105**	-0.058	-0.112**
mean twin pair birth weight)	(0.024)	(0.036)	(0.037)	(0.039)	(0.045)	(0.047)	(0.051)

Note: Column (1) present pooled grade three through eight results for the twin-FE model, with standard errors clustered at the individual level. Columns (2) to (7) present twin-FE estimates separately for each of the 6 grades. Each coefficient estimate comes from a separate regression (except for the last row where there are two coefficients from the same regression reported). Sample sizes and models are identical to these estimated in columns (2) and (3) to (8) in table 2 but the variable of interest is substituted. For the sake of clarity we carry over the main estimates from table 2 to the first row in this table. The second row presents the baseline model for the sample of twin pairs where both twins are above 2500g. The third row presents the baseline model for the sample of twin pairs where both twins are below 2500g. The fourth row presents the baseline model for the sample of twin pairs where both twins have birth weight between 1500g and 2499g. The fifth row presents the baseline model for the sample of twin pairs where both twins have birth weight below 1500g. The sixth row substitutes ln(birth weight) with birth weight measured in 1000g. The seventh row substitutes ln(birth weight) by birth weight in grams as the firs variable and the interaction between birth weight in grams and the difference of birth weight in grams and mean twin pair birth weight in grams as the second variable.

Table 5. Results by school quality measures

School quality measure	Sample	(1) % population	(2) Mean test score	(3) Mean (SD) birth weight	(4) Pooled twin FE estimate	(5) p-value of difference
(1) Awarded grade	A	48.6	0.277	2437 (559)	0.407*** (0.033)	0.204
	B	28.8	-0.095	2409 (570)	0.497*** (0.055)	
	C & D & F	22.6	-0.400	2375 (578)	0.455*** (0.062)	
(2) Average proficiency	Below median	39.7	-0.340	2381 (580)	0.436*** (0.048)	0.831
	Above median	60.3	0.297	2442 (555)	0.425*** (0.033)	
(3) Growth in proficiency	Below median	49.8	0.044	2420 (565)	0.449*** (0.036)	0.649
	Above median	50.2	0.098	2421 (564)	0.433*** (0.035)	

Note: Descriptive statistics for each group are reported in columns (1) to (2). Column (1) presents the fraction for each group within total population of twin pairs used in the analysis (born in Florida between 1992 and 2002 and successfully matched to Florida public schools). Columns (2) and (3) present mean combined mathematics and reading test scores and mean (SD) of birth weight for each studied group respectively. Column (4) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Column (5) presents the joint significance test for the analyzed groups in fixed effects model from column (4). In the case of awarded grades since not all schools are awarded grades every year our sample consist of 124,380 individual observations used in models in column (4). In the case of average proficiency and growth in proficiency we use 126,502 individual observations in models in column (4). The discrepancy between the samples in table 3 and table 5 is due to the fact that we do not have data on school quality for the universe of schools in every year in Florida (in particular average proficiency and growth cannot be calculated for a newly established school).

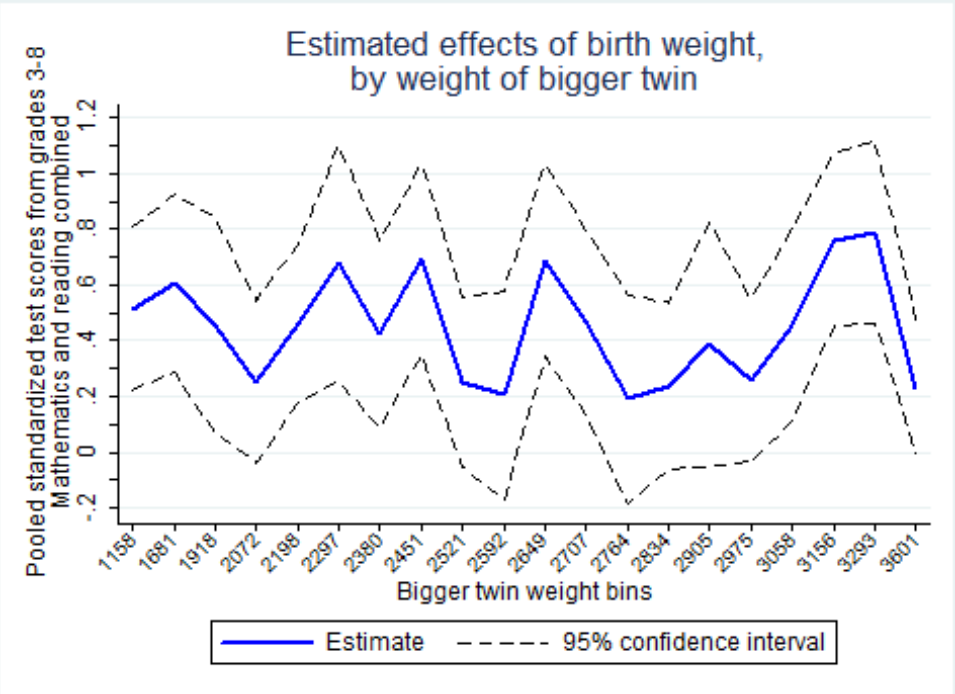
Table 6. Effects of birth weight on kindergarten readiness and comparison with FCAT scores

Panel	(1) N	(2) Kindergarten readiness measure (dichotomous)	(3) 3rd grade FCAT (discretized)	(4) Pooled panel FCAT (discretized)	(5) 3rd grade FCAT reading (discretized)	(6) 3rd grade FCAT math (discretized)
School readiness checklist only (98-01 KG cohorts)	8,938	0.067* (0.035)				
DIBELS only (06-08 KG cohorts)	6,696	0.115*** (0.043)				
Pooled SRC and DIBELS (KGR)	15,634	0.086*** (0.027)				
KGR & 3rd-8th grade panel	6,512	0.057 (0.040)	0.181*** (0.045)	0.146*** (0.024)	0.182*** (0.046)	0.149*** (0.045)
KGR & 3rd-5th grade panel	9,198	0.060* (0.033)	0.178*** (0.038)	0.167*** (0.022)	0.179*** (0.038)	0.166*** (0.037)
KGR & 3rd grade panel	13,718	0.093*** (0.029)	0.159*** (0.031)	0.159*** (0.031)	0.161*** (0.031)	0.138*** (0.031)
SRC & 3rd grade panel	7,824	0.060 (0.037)	0.163*** (0.040)	0.163*** (0.040)	0.098** (0.041)	0.103** (0.045)
DIBELS & 3rd grade panel	5,894	0.139*** (0.046)	0.118** (0.050)	0.118** (0.050)	0.101** (0.049)	0.103** (0.051)

Note: The first three rows present the estimated effects of ln(birth weight) on three kindergarten readiness measures. All models additionally control for infant gender, within-twin pair birth order and twin fixed effects. In each case the sample includes all twin pairs where both twins were assessed. The next three rows limit the sample to those with both a kindergarten readiness measure and, in turn, test scores observed between grades 3-8, grades 3-5, and grade 3. KGR refers to SRC and/or DIBELS. The discretized FCAT scores are created by assigning a value of 1 to the top 83 percent of the FCAT score distribution and 0 otherwise, in order to make the results directly comparable to those where the dependent variables are dichotomous kindergarten readiness indicators. The final two rows limit the sample to those with the SRC and 3rd grade score, and then the DIBELS and 3rd grade score, respectively. Standard errors are heteroskedasticity robust in cases in which there is one observation per individual and clustered to the student level in the cases in which there are multiple observations per individual. Sample sizes vary slightly from regression to regression depending on whether reading, math, or both scores are the dependent variable.

APPENDIX
FIGURES

Figure A1. Estimated effects of birth weight, by weight of bigger twin



Note: Figure A1 plots coefficients and 95% confidence intervals from a twin FE regression where the dependent variable is the mean of pooled grades three to eight combined mathematics and reading test scores for each individual and the independent variables are 20 interactions corresponding to the product of log birth weight with indicators for 20 bins reflecting heavier twin percentiled birth weight. The regression additionally controls for infant gender and birth order within-twin pair. Heteroskedasticity robust standard errors are used to calculate the 95% confidence interval. Numbers on the x-axis correspond to the mean birth weight in each bin of heavier twin birth weight.

TABLES

Table A1. Birth weight difference and test scores across imputed grades and groups: coefficients on log birth weight

Sample		(1) Pooled	(2)	(3)	(4)	(5)	(6)	(7)
			Imputed grade					
			3	4	5	6	7	8
Total sample		0.441***	0.442***	0.526***	0.430***	0.426***	0.386***	0.373***
		(0.029)	(0.043)	(0.045)	(0.047)	(0.053)	(0.056)	(0.061)
(1) Children gender composition	Same sex	0.447***	0.460***	0.527***	0.406***	0.464***	0.394***	0.363***
		(0.032)	(0.049)	(0.053)	(0.053)	(0.059)	(0.062)	(0.066)
	Opposite sex	0.427***	0.398***	0.524***	0.486***	0.335***	0.365***	0.395***
		(0.062)	(0.086)	(0.088)	(0.097)	(0.112)	(0.122)	(0.135)
(2) Gender	Boys	0.449***	0.471***	0.567***	0.379***	0.475***	0.410***	0.298**
		(0.052)	(0.089)	(0.096)	(0.097)	(0.114)	(0.115)	(0.121)
	Girls	0.444***	0.449***	0.488***	0.434***	0.453***	0.378***	0.428***
		(0.040)	(0.081)	(0.084)	(0.081)	(0.085)	(0.094)	(0.101)
(3) Same-sex composition	Girl-Girl	0.444***	0.449***	0.488***	0.434***	0.453***	0.378***	0.428***
		(0.039)	(0.067)	(0.070)	(0.068)	(0.071)	(0.078)	(0.085)
	Boy-Boy	0.449***	0.471***	0.567***	0.379***	0.475***	0.410***	0.298***
		(0.051)	(0.073)	(0.079)	(0.080)	(0.094)	(0.096)	(0.101)
(4) Maternal race (N=14 357)	White	0.466***	0.504***	0.546***	0.440***	0.419***	0.417***	0.389***
		(0.034)	(0.051)	(0.054)	(0.054)	(0.060)	(0.065)	(0.068)
	Black	0.381***	0.291***	0.476***	0.412***	0.447***	0.300**	0.341**
		(0.061)	(0.087)	(0.090)	(0.098)	(0.118)	(0.118)	(0.137)
(5) Maternal ethnicity	Non-Hispanic	0.434***	0.442***	0.518***	0.440***	0.395***	0.376***	0.358***
		(0.033)	(0.049)	(0.052)	(0.053)	(0.060)	(0.064)	(0.070)
	Hispanic	0.478***	0.442***	0.567***	0.390***	0.576***	0.432***	0.439***
		(0.059)	(0.094)	(0.092)	(0.103)	(0.110)	(0.115)	(0.125)
(6) Maternal immigration history	Non-immigrant	0.440***	0.469***	0.518***	0.439***	0.408***	0.367***	0.348***
		(0.033)	(0.049)	(0.052)	(0.053)	(0.061)	(0.065)	(0.070)
	Immigrant	0.449***	0.323***	0.563***	0.394***	0.510***	0.470***	0.478***
		(0.058)	(0.090)	(0.090)	(0.095)	(0.105)	(0.111)	(0.122)
(7) Maternal education	< 12	0.359***	0.249**	0.484***	0.431***	0.257**	0.369**	0.342**
		(0.070)	(0.110)	(0.125)	(0.121)	(0.128)	(0.145)	(0.152)
	<12; 15>	0.434***	0.466***	0.493***	0.410***	0.443***	0.365***	0.365***
		(0.038)	(0.055)	(0.055)	(0.060)	(0.070)	(0.070)	(0.078)
	> 15	0.529***	0.517***	0.656***	0.493***	0.503***	0.477***	0.433***
		(0.059)	(0.089)	(0.098)	(0.096)	(0.099)	(0.123)	(0.129)
(8) Zip code median income (N=11 868)	Bottom	0.389***	0.428***	0.445***	0.310***	0.328***	0.399***	0.396***
		(0.057)	(0.083)	(0.085)	(0.091)	(0.110)	(0.110)	(0.133)
	Middle	0.457***	0.409***	0.534***	0.491***	0.504***	0.387***	0.338***
		(0.054)	(0.081)	(0.089)	(0.086)	(0.101)	(0.116)	(0.126)
	Top	0.446***	0.507***	0.550***	0.383***	0.376***	0.320***	0.413***
		(0.059)	(0.085)	(0.088)	(0.098)	(0.108)	(0.121)	(0.144)
(9) Maternal marital status (N=14 583)	Non-married	0.362***	0.336***	0.402***	0.413***	0.376***	0.363***	0.218*
		(0.057)	(0.083)	(0.085)	(0.090)	(0.112)	(0.113)	(0.115)
	Married	0.485***	0.497***	0.588***	0.446***	0.454***	0.400***	0.458***
		(0.033)	(0.050)	(0.053)	(0.055)	(0.058)	(0.064)	(0.072)
(10) Maternal age at birth of children	<= 21	0.372***	0.371***	0.411***	0.495***	0.237	0.399**	0.233
		(0.086)	(0.116)	(0.130)	(0.136)	(0.172)	(0.168)	(0.177)
	<22; 29>	0.443***	0.417***	0.509***	0.374***	0.534***	0.417***	0.385***
		(0.044)	(0.067)	(0.066)	(0.071)	(0.081)	(0.085)	(0.093)
	<30; 35>	0.483***	0.466***	0.585***	0.496***	0.465***	0.388***	0.426***
		(0.052)	(0.080)	(0.081)	(0.085)	(0.090)	(0.101)	(0.113)
	>= 36	0.413***	0.529***	0.570***	0.393***	0.182	0.270*	0.354**
		(0.078)	(0.114)	(0.135)	(0.124)	(0.134)	(0.155)	(0.155)

Note: Column (1) present pooled grade three through eight results for twin-FE model. Columns (3) to (8) present twin-FE estimates separately for each of the 6 grades. Models are the same as used in columns (2) and (3) to (8) in table 2. Sample size is 127 156 individual observations in pooled regressions in column (1) except for race, marital status and mean zip code income. In the case of race this discrepancy is caused by existence of other races with minor representation in Florida, In the case of income and marital status we do not have complete data for all mothers and residential locations. In all these cases the modified sample sizes are given. Each coefficient comes from a separate regression.

Table A2. Sensitivity to model specification: Birth weight as a second order polynomial

Sample	(1) Pooled	(2)	(3)	(4) Imputed grade	(5)	(6)	(7)
		3	4	5	6	7	8
Birth weight	0.450***	0.477***	0.487***	0.466***	0.446***	0.317***	0.441***
	(0.063)	(0.094)	(0.098)	(0.100)	(0.118)	(0.122)	(0.133)
Birth weight^2	-0.053***	-0.058***	-0.053***	-0.057***	-0.053**	-0.029	-0.056**
	(0.012)	(0.018)	(0.019)	(0.019)	(0.023)	(0.023)	(0.025)

Note: Column (1) present pooled grade three through eight results for twin-FE model. Columns (2) to (7) present twin-FE estimates separately for each of the 6 grades. Both coefficients comes from the same regression. Sample sizes and models are identical to these estimated in columns (2) and (3) to (8) in table 2 but the variable of interest (ln(birth weight)) is substituted by birth weight in grams and its square.