Institute for Policy Research
Northwestern University
Working Paper Series

WP-13-08

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

David Figlio
Orrington Lunt Professor of Education and Social Policy and of Economics
Director and Faculty Fellow, Institute for Policy Research
Northwestern University

Jonathan Guryan
Associate Professor of Human Development and Social Policy
Faculty Fellow, Institute for Policy Research
Northwestern University

Krzysztof Karbownik
Visiting Scholar, Institute for Policy Research
Northwestern University

Jeffrey Roth
Research Professor of Pediatrics
College of Medicine
University of Florida

Version: October 2014

DRAFT
Please do not quote or distribute without permission.

Abstract

This working paper makes use of a new data resource—merged birth and school records for all children born in Florida from 1992 to 2002—to study the effects of birth weight on cognitive development from kindergarten through schooling. Using twin fixed effects models, the researchers find that the effects of birth weight on cognitive development are essentially constant through the school career, that these effects are very similar across a wide range of family backgrounds, and that they are invariant to measures of school quality. They conclude that the effects of poor neonatal health on adult outcomes are therefore set very early.

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, Chile, China, Norway, and the United States, have made use of twin comparisons to show that the heavier twin of 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 is important for a variety of policy reasons to know how poor neonatal health affects child development, and whether there are public policies that might act to remediate the negative relationship between early poor health and later-life outcomes. Knowing this relationship can also be useful in helping to understand whether favorable health at birth can shield children against adverse shocks, policy or otherwise. However, we know very little to date about whether the effects of poor neonatal health on cognitive development vary at different ages (say, at kindergarten entry versus third grade versus eight grade), and no existing study identifies whether public policies such as school quality could help to mitigate the effects of poor neonatal health on cognitive development. We also know very little about whether these effects vary heterogeneously across different demographic or socio-economic groups, or whether early neonatal health and parental inputs are complements or substitutes. While we have strong evidence

[1] Examples of influential previous research include Behrman and Rosenzweig (2004) on schooling and wages, Almond, Chay, and Lee (2005) and Conley, Strully, and Bennett (2003) on neonatal outcomes and hospital costs, and Royer (2009) on next generation birth weight, neonatal outcomes and educational attainment, for the United States; Black, Devereux, and Salvanes (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 (2014) on educational attainment, wages and weight for height, for China; and Torche and Echevarria (2011) on fourth-grade mathematics test scores, for Chile. In a current working paper, Bharadwaj, Eberhard, and Neilson (2013) study fourth-grade test scores and grades in school, also in Chile.

from twin comparison studies that poor initial health conveys a disadvantage in adulthood, we have little information about the potential roles for policy interventions in ameliorating this disadvantage during childhood.

The principal 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.[2]

Another gap in the adult-outcomes literature is that the subjects of that literature 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.[3] There have been no studies linking neonatal health to either educational or later outcomes in a highly developed country context using very recent birth cohorts.[4]

We make use of a major new data source that can help 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

[2] Exceptions include Bharadwaj, Eberhard, and Neilson 2013; Torche and Echevarria, 2011; and Rosenzweig and Zhang, 2009, which examine this relationship in developing countries with less access to advances in medical technology that have reduced the lower end of viable birth weights, and in settings that lack the socioeconomic and ethnic diversity that is present in the data from Florida used in this paper. Another alternative data source is the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) of children born in the United States in 2001 which oversamples twins. However, the ECLS-B 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.

[3] One example of the temporal differences in neonatology is that, whereas 50 years ago the threshold for infant viability was around 1500 grams, today the threshold for viability in developed countries is as low as 500 grams or even lower (Lau et al., 2013). Thus, 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.

[4] The potential benefits of using more current data from a highly developed country become apparent when we compare the mean birth weight amongst twins in our study of children born after 1992 (2410 grams) to those from previous studies of twins from highly developed countries born in the 1930s through the 1970s (which range from 2517 to 2598 grams, depending on the cohort and country) and those from the late 1990s in Chile (2500 grams).

around 17 million compares to Norway, Denmark, and Sweden combined) and heterogeneous (nearly half of mothers are racial or ethnic minorities, and nearly one-quarter of mothers were born outside the United States). In addition, Florida has some of the strongest education data systems in the United States, and Florida has been testing children annually from third through tenth grade for well over a decade. With these new data, we follow over 1.3 million singletons and nearly 15,000 pairs of twins from birth through middle school to study the relationship between birth weight and cognitive development.

We find that neonatal health, as measured by birth weight, affects cognitive development in childhood, and that this relationship is remarkably consistent across subgroups from a wide range of family socio-economic status.[5] We observe this relationship for twin-pair comparisons, sibling-pair comparisons, and singletons, and while the magnitudes of these comparisons differ somewhat, they provide reasonable bounds of the likely effects of neonatal health on children's cognitive outcomes.

Comparing across a range of demographic and socio-economic dimensions allows us to address both the stability of results across background and the degree to which parental inputs and early health are complements or substitutes. Understanding this complementarity is important because it provides a window into the mechanisms by which neonatal health and parental resources and behavior contribute to human capital development. Whether parental inputs and neonatal health are complements or substitutes also has important implications

[5] We are certainly not the first paper to conduct heterogeneity analyses of families with twins. Black, Devereux and Salvanes (2007) mention that they investigated sample splits by income and education and find no significant differences, but do not report their subgroup-specific findings, making it impossible to address the question of whether parental inputs and early health are complements or substitutes. Oreopoulos et al. (2008) report results broken down by birth weight group, gestational length, and APGAR scores, but not by different socio-economic groups. Johnson and Schoeni (2011) report results by parental age and the presence of health insurance, which could reflect a variety of factors other than the key questions that we are interested in studying. Bharadwaj, Eberhard, and Neilson's (2013) working paper and Torche and Echevarria (2011) split their analyses by maternal education – but the developing Chilean context at the time means that Bharadwaj, Eberhard, and Neilson (2013) only split by high school and over versus middle school or lower education.

theoretically for understanding the distributional effects of investments in infant health, and for guiding the targeting of policies intended to reduce inequalities by improving early life health (e.g. consider the role complementarities play in the models of human capital accumulation of Cuhna et al. (2006), Cunha and Heckman (2007), and Conti and Heckman (2010)). We find evidence that the effects of birth weight on student outcomes are stronger for higher-SES families than for lower-SES families, suggesting that neonatal health and parental inputs are at least to some degree complements. Such complementarity could be driven by parents with more resources investing more in children with better neonatal health, or could be the result of parents making equal investments but those investments by more educated higher-SES parents being relatively more or less effective at building the human capital of children born with better initial health.

Importantly, ours is the first study to explore the interaction between schooling factors and the relationship between birth weight and children's cognitive development. Once children reach school age, they spend considerably more time with adults who are not their parents than they did before school age. Schooling is the most natural place where public policy can play a role in promoting cognitive development amongst children in non-familial settings. We seek to understand the degree to which school quality can help to overcome disadvantages associated with poor neonatal health. We find that the relationship between birth weight and cognitive outcomes is invariant to a variety of measures of school quality, suggesting that while high-quality schools have the potential to improve the outcomes of all children, they do not reduce the gaps generated by poor neonatal health.

I. A new data source

A. Description of the data set and match diagnostics

We make use of matched data for all children born in Florida between 1992 and 2002 and educated in a Florida public school between 1996 and 2012. For the purposes of this study, Florida's education and health agencies matched children along three dimensions: first and last names, exact date of birth, and social security number, with a small degree of fuzziness permitted in the match. Common variables excluded from the match were used as checks of match quality. These checks confirm that the matches are very clean: In the overall 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 were 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, 80.7 percent of all children born in Florida, and 79.5 percent of all twins born in Florida, were matched to school records using the match protocols.

In order to judge the quality of the match, we compare the 80.7 percent rate to population statistics from the American Community Surveys and Census of Population from 2000 through 2009.[6] 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

[6] The benefit of non-name unique match identifiers in Florida becomes apparent when we compare our 80.7 percent match rate to the match rate in North Carolina, the only other state where, to our knowledge, researchers are making use of matched birth-school data today. The cleanest North Carolina match rate, which relies on children being matched by name, date of birth, and county, is just over 50 percent, and when the match is made less exactly, just on name and date of birth, the match rate in North Carolina is between 60 and 65 percent, depending on subgroup (Ladd, Muschkin, and Dodge 2012).

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.[7] 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

The set of Florida-born children attending Florida public schools differs fundamentally from the set of all Florida-born children. It is important to note that twins differ from singletons in important ways. Twins have a lower mean gestational age and a lower mean birth weight than singletons; they have older and more educated mothers, as well as mothers who are more likely to be married (Antsaklis et al., 2013). We discuss issues of external validity in the conclusion.

[Insert Table 1 Here]

Table 1 presents some evidence regarding the overall representativeness of the population of children matched to schools and the 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

[7] The 80.9 percent 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.

born in Florida; the second reflects the population of children matched to Florida public school records; the third represents the set of children with a third grade test score; and the fourth reflects the set of twins born in Florida who have a third grade test score. (Children in these last two columns also must fulfill the other data requirements, such as non-missing core control variables, for inclusion in the study.) 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 (17.1 percent college graduate versus 20.1 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 9.3 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. Third-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.[8] However, our findings of estimated relationships between birth weight and test scores that are remarkably

[8] 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.

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. Mothers of twins are quite different from the overall population: 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 (23.1 percent college graduate versus 15.8 percent in the overall population of test-takers, and 15.5 percent high school dropout versus 23.3 percent of all test-takers) and considerably older (13.6 percent aged 36 or above versus 9.2 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.)[9] Therefore, the decision to focus on twin comparisons to promote increased internal validity brings with it some cost in terms of external validity. In this paper, we therefore present evidence on the relationship between birth weight and cognitive development both in the case of twin comparisons – where internal validity is greatest – as well as the case of singletons – where external validity is greatest. Our general patterns of results are quite similar across both cases.

C. Birth weight distributions

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 substantial. In Florida, the average discordance in twins' birth weight is 284 grams (0.63

[9] 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.

pounds), or 11.8 percent of the average twin's birth weight of 2410 grams.[10] Figure 1 shows that the distribution of discordance for all twins is virtually identical to the distribution of discordance for twins matched to test scores. 51.4 percent of twin pairs have birth weight discordance over 200 grams, and 16.8 percent have birth weight discordance over 500 grams. Forty five 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.[11]

[Insert Figure 1 Here]

[Insert Figure 2 Here]

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 (2410 grams) is 27.9 percent smaller than the mean singleton birth weight of 3342 grams. 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. 53.2 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.

[10] Blickstein and Kalish (2003) provide an overview of the literature on growth restriction explanations for birth weight discordance. In addition, there are some medical reasons that might lead to birth weight discordance; for example, Kent et al. (2011) find that noncentral placental cord insertion leads to birth weight discordance in some pregnancies. Breathnach and Malone (2012) survey the literature on fetal growth disorders in twin gestations.

[11] 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.

II. Empirical framework

Our empirical framework largely follows what has become standard in the literature. For our twins' analysis, we estimate twin fixed effect models in which the regressor of interest is the natural logarithm of birth weight.[12] Following Almond, Chay and Lee (ACL, 2005) and Black, Devereaux and Salvanes (BDS, 2007), let

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

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 child-specific determinants of the outcome (in the case of twins, child gender and within-twin-pair birth order), ϕ denotes unobservable determinants of the outcome that are specific to the mother and birth, and ε is an error term. We also estimate singleton-specific analyses in which we control for a wide range of maternal characteristics, as well as (in some specifications) gestational length, to make as apples-to-apples comparisons with the twin specifications as possible. Our results are invariant to whether or not we condition on geography.

Our 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.[13] For ease of presentation, we average standardized reading and mathematics FCAT scores for our dependent variable, but our results are qualitatively similar for reading and mathematics, and the test-specific results are

[12] We follow an analogous approach regarding sibling comparisons.

[13] 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.

available on request.[14] 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 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. We also present evidence from singleton births that, while they lack the internal validity of the twin comparisons, allow us to show the relationships between gestation length, birth weight, and cognitive skills in the overall population of children.

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, the absence of one twin's test score could present a source of bias. A related concern relates to the fact that we only observe education records for individuals born in Florida who remained

[14] In the main twins regression specification, 99.5 percent of observations have both math and reading scores, 0.2 percent have only math and 0.3 percent have only reading.

in Florida, attended Florida public schools and took the FCAT. Various tests reported in detail in Figlio et al. (2013) suggest that in practice the selection bias resulting from either of these sources is likely to be minimal. For example, the likelihood of leaving the sample between 3rd grade and 4th or 5th grade is uncorrelated with whether the twin is the heavier or lighter of the pair, and only slightly more likely for the lighter twin in grades 6-8. The relative number of missing twins is too small to make a meaningful difference in the estimates even in these later grades. Furthermore, estimates in which we impute very low or very high test scores for missing twins yield almost identical results as those reported in the main specifications.

III. Preliminary results – heavier versus lighter twins

To fix ideas before presenting the main regression results, figure 3 shows the average within twin pair difference in average math and reading test score between the higher birth weight twin and the lower birth weight twin for grades three through eight.[15] Within twin pairs, on average the heavier twin scores about five percent of a standard deviation higher than the lighter 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.[16] The results imply that neonatal 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 middle school.

[Insert Figure 3 Here]

[15] The same patterns for math and reading separately are in figures A1 and A2 in an online appendix.

[16] 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.

[Insert Figure 4 Here]

Figure 4 breaks down this mean difference by quartile of twin birth weight discordance[17]; the bottom and top quartiles average 2.5 and 23.9 percent discordance, respectively. Two facts are apparent from this figure: First, the relationship between relative birth weight and relative test scores within twin pairs is roughly flat as children age. Second, the higher degree of birth weight discordance, the larger test score gap between the larger and the smaller twin. Figure A3 in an online appendix shows that the positive relationship between birth weight discordance and test score differences is present and clear when we break down the twin pairs or sibling pairs into fine discordance bins (one for each percentage point, and a final bin for twin pairs with greater than twenty percent discordance), with the slope of the relationship modestly flatter for sibling pairs than it is for twin pairs. These findings foreshadow the main findings of this paper.

IV. Main results

A. Pooled results for full sample

We now turn to our main regression results. The basic regression model is an OLS 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 score averaged between reading and math[18], 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

[17] We limit this analysis to same-sex twins to ensure that the differences in discordance are not due to well-documented differences in birth weight between boys and girls.

[18] See Figlio et al. (2013) for separate findings for reading and mathematics.

other results that pool test scores across all six grades. In the pooled regressions, standard errors are clustered at the individual level (for singletons) and twin-pair level (for twins) to account for the fact that each individual has up to six observations, one for each grade in which he or she was tested.[19]

[Insert Figure 5 Here]

The non-parametric plots of the relationship between test scores and birth weight reported in figure 5 present evidence supportive of the log birth weight specification that we employ, as there appears to be a concave relationship between birth weight and test scores. The figure shows two series, each derived from a test score regression that pools grades 3-8 and both math and reading scores. Each series plots the coefficients from a set of 36 dummy variables corresponding to 100 gram-wide birth weight bins. The bins range from a low of 501-600g to a high of 4,001-4,100 grams. In both regressions, the left-out group is below 501 grams. As was observed in similar sets of plots by ACL and BDS, the shape of the relationship between test scores and birth weight is similar whether or not we condition on twin-pair fixed effects.

[Insert Table 2 Here]

The main result, an estimated coefficient of 0.443 presented in column 2 of table 2, implies that a ten percent increase in birth weight is associated with just under one-twentieth of one standard deviation increase in test scores in grades three through eight.[20] The coefficient is precisely estimated, with a *t*-statistic of

[19] An earlier version of this paper (Figlio et al., 2013) clusters standard errors for twins at the individual level. The level of clustering (individual versus twin pair) has no substantive effect on our findings. In grade-by-grade singleton models with one observation per child, we estimate robust standard errors.

[20] We also find that birth weight is associated with a modest but strongly statistically significant increase in a child's grade in school at any given age. In the twin fixed effects model, a ten percent increase in birth weight is associated with

over 10. The fixed effects result is modestly larger than, but close to, the equivalent OLS coefficient of 0.285 reported in the first column of table 2.[21]

To put the magnitude of these coefficients 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. Similarly, Royer (2011) estimates that a 1000 gram increase in birth weight is associated with an extra 0.16 years of schooling. Using the online analysis tool of the High School & Beyond data set, which longitudinally follows a cohort in the middle of Royer's sample, we estimate that a one standard-deviation increase in 10th grade test scores is associated with 0.84 additional years of completed education.[22] Combining this with our finding that a 1000 gram increase in birth weight is associated with a 0.187 standard deviation increase in test scores, our results imply a 1000 gram increase in birth weight is associated with 0.156 additional years of schooling, almost exactly in line with Royer's findings.

[Insert Figure 6 Here]

Our estimate of the effect of neonatal health on cognitive development is reasonably large in these terms, but it is worth comparing to other important correlates of student achievement. Figure 6 shows that the difference in test scores

---

just under 1/100 higher grade for any given age; the estimated coefficient on log birth weight when the dependent variable is grade for age is 0.083 with a standard error of 0.019.

[21] We concentrate on birth weight because there is greater variation in birth weight than in other measures of neonatal health. That said, we find positive, statistically significant relationships between APGAR scores and test scores. For instance, in a pooled twin fixed effects model, a one unit increase in one minute APGAR scores is associated with 0.8 percent of a standard deviation higher average reading and math scores.

[22] We weighted the individuals in the High School & Beyond data by their base year replicate weights. For the sake of this analysis, we define high school dropouts as having 10 years of education, GED recipients as having 11, high school graduates as having 12, certificate recipients as having 13, associates recipients as having 14, bachelors recipients as having 16, masters or professional degree recipients as having 18, and doctorate recipients as having 19 years of education.

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 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 twin of a college educated mother than the heavier twin of 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.[23]

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 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 develop, or does the effect appear by age 9 and remain constant through the upper elementary and middle grades?

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. The table shows that the twin fixed estimate of the effect of birth weight on cognitive achievement is already 0.444 by the third grade, and that the grade-specific estimated effect remains fairly stable from third through eighth grade, ranging

[23] We do not mean to suggest that our 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).

from 0.376 to 0.526. The *F*-test that the grade-level estimated effects are identical is rejected at a moderate level of statistical significance (p=0.069). However, there is no evidence that this effect follows a substantial systematic pattern as children progress through school; 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 log birth weight. These results suggest that whatever effect early health at birth has on cognitive development occurs largely by age 9, and remains fairly constant throughout the preadolescent and adolescent years.

In a previous version of this paper (Figlio et al., 2013), we look further back, to the beginning of formal schooling.[24] In various years between 1998 and 2008, Florida performed universal kindergarten readiness screening. 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.[25] 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. Making use of twin comparisons in a linear

[24] There is some reason to believe that the effects of early health deficits may differ between the start of kindergarten and the end of third grade. 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 how the effect of neonatal health on cognitive development changes in the early schooling years.

[25] For more details about the structure and interpretation of DIBELS, see, e.g., Hoffman, Jenkins, and Dunlap (2009).

probability model[26], we observe 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 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.[27] All estimates are statistically distinct from zero at conventional levels. These results suggest that the effect of neonatal health on cognitive development is present by age 5.

C. Role of genetic differences between twins

For some policy questions, it might be important to isolate the impact of factors that change intrauterine growth while holding genetics constant. A potential weakness of our data is that they do not include the zygosity of the twins. However, we can look at same-sex versus different-sex twins: 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 weight and test scores among different-sex twin pairs. As can be seen in second and third rows of table 2, the estimated effect of birth weight is extremely similar for same-sex twins (0.452) and different-sex twins (0.421), suggesting that the estimated relationship is within the same general range regardless of zygosity. Our finding is consistent with results reported in BDS, who find no significant difference in the effect of birth weight on adult earnings between same-sex and opposite-sex twins, nor do they find a significant

[26] The pattern of results and statistical significance is extremely similar when we instead estimate conditional logit models.

[27] In Figlio et al. (2013) we go into detail about the metrics one can employ to directly compare the dichotomous kindergarten readiness assessments to later continuous test scores.

difference in the estimated effect of birth weight on earnings for monozygotic twins and dizygotic same-sex twins in their sample with available zygosity information.

D. Parallel results for singletons

As mentioned above, our emphasis (and the prevailing emphasis in the literature) on using twin comparisons to improve internal validity comes at a cost in terms of external validity. Twins have older and more educated mothers, and weigh considerably less on average at birth than singletons. In addition, there could be some unmeasured factor (e.g., a factor associated with in-utero fetal competition) associated with both birth weight and cognitive skills that could compromise our ability to draw causal inferences about the effects of neonatal health on later test scores in twin comparison studies. For these reasons, it is valuable to gauge the degree to which the estimated relationships for singletons compare with the findings for twins. In our singletons regressions, we further control for a set of background characteristics - gender, month and year of birth dummies, marital and immigrant status, race and ethnicity, three dummies for maternal education, and dummies for age and number of prior births.

The fourth row of table 2 presents OLS findings for singletons. Two features are apparent: First, the relationship between log birth weight and test scores is roughly constant as children grow older, just as it was in the case of twins. Furthermore, the OLS coefficient for singletons in the pooled model (0.285) is identical to the comparable OLS coefficient for twins (0.285). This similarity provides the first piece of evidence about the potential external validity of our twin results.

Recall that our twin fixed effects relationship is larger than our twin OLS relationship. One possible reason for this difference is that the twin fixed effects

relationship effectively conditions on gestational length. In the fifth row of table 2 we condition on gestational length for singletons, and find an OLS coefficient that is somewhat larger than was the case without controlling for gestational length. A comparison of the results may indicate that the rate of intrauterine growth matters for cognitive development, above and beyond the effect of measured birth weight.

Singletons include some infants whose birth weight is high enough that it likely indicates an underlying poor maternal health condition such as gestational diabetes, whereas it is rare for a twin to have a birth weight in this high range. When we further limit the singletons analysis to the range between 847 and 3600 grams, the 1st and 99th percentiles of the twin birth weight distribution, we estimate the OLS relationship between log birth weight and pooled test scores, conditional on gestational length, to be 0.421, extremely similar to the twin fixed effects finding of 0.443. In sum, the closer we get to shaping the singletons OLS analysis to be parallel to the twin fixed effects analysis, the closer the two results become. In addition, as can be seen in the seventh row of the table, when we look just at the relationship between weeks of gestation and standardized test scores, we observe that each week of gestation is associated with just over one percent of a standard deviation increase in test scores.

In a set of counties representing 56 percent of the population of the state of Florida, we are able to also control for family fixed effects in the singletons analysis. The results of this sibling analysis are presented in the eighth through tenth rows of table 2. The estimated effects of birth weight on test scores in the sibling comparisons tend to be around three-quarters of the magnitude of the twin fixed effects estimates, but remain in the same ballpark. The differences in magnitudes are due to the differences between the sibling comparisons and the twin comparisons, and not the fact that we observe siblings in a subset of the state, as can be seen when we consider the OLS coefficients in the sibling subpopulation to the overall singletons population. The OLS coefficient on log

birth weight is 0.277 for siblings and 0.285 for singletons, and the coefficient on log birth weight conditional on gestation in the overlapping sample is 0.403 versus 0.421 for all singletons. We suspect that the modest differences between the twin fixed effects models and sibling fixed effects models are due to factors such as differential parental investments in siblings (Bharadwaj, Eberhard, and Neilson, 2013; Hsin, 2012) or direct spillovers between siblings (as we find in Black et al., 2014).

[Insert Figure 7 Here]

Since we find that the estimated coefficients on log birth weight are so similar when we condition on twin fixed effects or when we use the population of singletons with birth weights in the observed range of twins and condition on gestation length, a natural next step is to observe whether the distribution of these estimated effects are the same as well. In figure 7, we present the estimated marginal effects of log birth weight on different parts of the CDF of the test score distribution, broken down by half-standard-deviation increments, for twins, singletons, and siblings. This figure demonstrates that additional birth weight is especially strongly associated with moving children from the range of scores just below average to the range of scores just above average, and is less strongly related to test scores far away from the average score.

E. Heterogeneity of results by gender, maternal health, and background

The diversity of demographics in Florida combined with the size of the dataset 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. It is inherently interesting to learn whether the long-term effects of *in utero* conditions on cognitive development vary across demographic and socio-economic groups.

Moreover, examining this heterogeneity may shed light on the mechanisms by which neonatal health affects cognitive skills. If the factors of disadvantage – e.g., household income, wealth and parental education – are substitutes with neonatal health in the production of cognitive skills one should expect to see larger effects of birth weight on test scores for more disadvantaged groups. If they are complements with neonatal health, one should expect to see larger effects for more advantaged groups.

[Insert Table 3 Here]

Table 3 presents a wide range of heterogeneity findings. For the sake of clarity, in the table we report the results in which we pool test scores across all grades; in online appendix table A1 we report grade-by-grade results for all subgroups of the twins analysis. Furthermore due to space constraints, in the print appendix table A1 we report the group mean test score and birth weight for twins and singletons, respectively, in each subgroup. The first column in table 3 reports the mean and standard error of the estimated effect of birth weight on test scores in a twin fixed effects model. The second through fourth report the parallel findings for singletons: The estimated coefficient on log birth weight (column 2), log birth weight conditional on gestation length (column 3)[28], and gestation length (column 4), while the fifth through seventh columns perform the same analysis when we condition on sibling fixed effects.

As can be seen in the first panel of table 3, the results are very similar for boys and girls.[29] While boys are heavier than girls (4.4 percent for twins, 3.8 percent for singletons), the pooled twin fixed effects estimates for boys and girls are

[28] In the singleton and sibling specifications conditioning on gestational length, we also limit the range of birth weights to the approximate twins birth weight range, between 847 and 3600 grams.

[29] Rosenzweig and Zhang (2009) suggest that there could be important differences by gender in their study's setting. However, these differences may reflect cultural factors specific to the rural Chinese context.

virtually identical (0.454 and 0.449, respectively). The same is true when we make comparisons in either the singleton population or in the case of sibling fixed effects.

The second panel of table 3 stratifies births based on whether the mother has a medical history that potentially posed a problem for the pregnancy or delivery.[30] Around one-quarter of mothers have at least one of these risk factors. We observe that the pooled fixed effects estimates are very similar (0.422 for mothers with medical history, and 0.449 for mothers without medical history), as are the log birth weight coefficients for singletons (for instance, 0.372 for mothers with medical history, and 0.437 for mothers without medical history in the case where we condition on gestational length). These results indicate that maternal health at the time of labor and delivery does not appear to matter much in terms of the effects of birth weight on cognitive development.

The third through ninth panels of table 3 show estimates of the effect of birth weight on pooled third through eighth grade test scores separately by maternal race (panel 3), maternal ethnicity (panel 4), maternal immigrant status (panel 5), maternal education (panel 6), a proxy for family income – the zip code's median income as of the 2000 Census (panel 7), maternal marital status (panel 8), and maternal age at the time of the child's birth (panel 9). These factors represents a massive range of student advantage, with average group test scores among twins as low as -0.475 and as high as 0.663 (see print appendix table A1), reflecting gaps that are consistent with other studies of U.S. school children (e.g., Chay, Guryan, and Mazumder 2009). Strikingly, the twin fixed effects coefficient estimates are remarkably similar across this wide range of groups, with point estimates ranging between 0.358 and 0.523. The OLS coefficient estimates in the

[30] The specific medical history factors recorded on the birth record are anemia; cardiac disease; acute or chronic lung disease; diabetes; genital herpes; hydramnios/oligohydramnios; hemoglobinopathy; chronic hypertension; pregnancy-associated hypertension; eclampsia; incompetent cervix; previous infant over 4000 grams; previous preterm or small for gestational age infant; renal disease; RH sensitization; uterine bleeding; and other specified history factors.

singleton population range from 0.249 to 0.326, and the OLS coefficient estimates on birth weight conditional on gestation range between 0.344 and 0.490. Sibling fixed effects coefficients conditional on gestation range from 0.282 to 0.418. Taken together, these results indicate that the effects of birth weight on test scores are roughly the same for children from a wide range of different backgrounds.

F. Complementarity of neonatal health and parental inputs

A close look at the subgroup analysis can provide some evidence regarding the degree to which neonatal health and parental inputs are complements or substitutes. One might expect parents with more resources to be better able to remediate the effects of poor neonatal health. However, whether neonatal health and parental inputs are complementary is determined by whether parents with more resources are relatively more effective at building human capital for children of good versus poor neonatal health, which could happen either because parents with more resources invest more or because the investments they make have higher returns.[31] Learning whether parental resources and neonatal health are complementary provides a window into mechanisms by which parents and early health interact in the human capital development process.

[Insert Figure 8 Here]

To explore this question systematically, we pursue an approach similar to that employed by Hoynes, Miller, and Simon (2014) to study the relationship between the Earned Income Tax Credit (EITC) and rates of low birth weight for different groups broken down by their rate of EITC usage. In our case, we use maternal race, maternal ethnicity, maternal immigrant status, maternal marital status,

[31] See Guryan, Hurst and Kearney (2008) for evidence that more educated parents spend more time in parenting activities with their children, and for a discussion of how that could theoretically result from either a desire to invest more or from higher returns.

maternal age, maternal education, and neighborhood income to predict student test scores in order to construct an index of the family socio-economic status (SES), and then divide the students into ten mutually-exclusive groups; these groups range in mean predicted test scores from -0.701 to 0.809 in the twins population – a range greater than a full individual-level standard deviation of the test score distribution.[32] Figure 8 plots each group's estimated coefficient on log birth weight against the group's mean score. We explore the relationship between SES and the effect of birth weight on children's cognitive development in three different models – the twin fixed effects model, the sibling fixed effects model conditional on gestation and restricted to the population of singletons whose birth weights fall within the observed range of twin birth weights, and the comparable OLS model for singletons.

The figure demonstrates two important features of the heterogeneity of birth weight effects across a wide range of groups stratified by predicted test scores. First, the estimated effects of birth weight are all within the same general range between 0.30 and 0.67 in the twin fixed effects model, between 0.29 and 0.48 in the singletons OLS model, and between 0.24 and 0.45 in the sibling fixed effects model, and the estimated effects are both statistically and economically significant for every demographic and socio-economic group analyzed.[33] These magnitudes would imply that the effects on cognitive development could account for half to all of the long-term relationship between birth weight and earnings estimated by BDS.

[32] The groups range in mean test scores from -0.618 to 0.755 in the case of singletons and from -0.696 to 0.817 in the case of sibling fixed effects.

[33] We have also estimated specifications in which we interact log birth weight separately with the socioeconomic variables referenced in table 3. We then evaluated the marginal effect of log birth weight separately for every child in the population. The marginal effects in the case of the twin fixed effects specification ranged from 0.17 to 0.62. Online appendix figure A4 plots the estimated marginal effects of log birth weight for the full distribution of possibilities in this specification.

The second pattern the figure illustrates is 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 larger for relatively advantaged groups of children than they are for relatively disadvantaged groups of children. The slopes of the lines plotted in figure 8 are 0.132, with a standard error of 0.086, in the case of the twin fixed effects model, 0.136, with a standard error of 0.060, in the case of the sibling fixed effects model, and 0.083, with a standard error of 0.019, in the case of the singletons OLS model.[34] The three lines are similar in terms of both slope and intercept, and indeed, the twin fixed effects and sibling fixed effects lines are virtually parallel. It is highly unlikely that these results are driven by differential selection into the sample across groups, at least by birth weight. As an illustration, the difference in gaps in average birth weight between twin-pairs with test scores and those without test scores ranges from -47 grams to 82 grams, and follow no apparent pattern: The typical gap is just 3 grams for the bottom half of the SES distribution and 7 grams for the top half. Therefore, while by no means definitive, these patterns indicate that poor neonatal health may disproportionately affect children growing up in high socio-economic status families, and are suggestive that neonatal health and parental resources are to some degree complementary.[35]

[34] We estimate the standard errors of the slopes of these lines by bootstrapping. We randomly drew twin pairs (sibling pairs or singletons) with replacement to generate a sample of the same size as our analysis sample. We then used this sample to predict test scores and to separate the bootstrapped sample into ten deciles based on predicted test scores. Next, we estimated twin fixed effects (sibling fixed effects or singleton) models for each of the ten deciles. For both twins, siblings and singletons, we ran 1000 replications of these 10-observation regressions and calculated the standard deviation from these slopes for our bootstrapped standard errors.

[35] Children in higher-scoring subgroups – who tend to have 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 8 – at least for the twins population - is due at least in part to differential patterns of IVF/ART. This association could be especially important in a population of twins, given that Bitler (2008) demonstrates that requiring health insurance plans to cover use of IVF/ART substantially increases the likelihood that a mother will have twins, and these new 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. Bitler uses this age breakdown to proxy for IVF/ART likelihood. Next, we conduct the identical analysis for twins who were the first children

V. Effect variation across the birth weight distribution and by discordance levels

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. Public health officials and medical practitioners frequently direct attention on the thresholds of 1500g and 2500g, the conventional delimiters of very low birth weight and low birth weight, respectively. 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 effects were only present in comparisons between moderate and high birth weight infants.

We have already presented non-parametric evidence (figure 5) 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 and explore how the marginal effect of poor neonatal health varies across the distribution of birth weight and with birth weight discordance. First we estimate models that allow the marginal effect of log birth weight to vary across different bins of the birth weight distribution. As seen in figure 9, which presents separate twin fixed effects coefficients for 20 equally-sized bins, based on the

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. We do not find evidence that these slopes differ appreciably across these groups of mothers. 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.

lighter-born twin's birth weight,[36] we observe no systematic relationship between the marginal effect of log 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.[37]

[Insert Figure 9 Here]

[Insert Figure 10 Here]

We next explore whether the relationship between birth weight and test scores varies by birth weight discordance in figure 10. We divide twins into 20 bins by birth weight discordance, excluding the twin pairs that are very close in weight (<150g difference).[38] As can be seen in the figure, the estimated relationship between log birth weight and test scores is qualitatively similar across a wide range of discordance.

Given the salience in the medical and public health literature of specific birth weight thresholds (1500g and 2500g), we next explore whether the estimated effects of log birth weight in twin fixed effects models differs systematically above and below 2500g. Rows 2 through 5 of online appendix table A2 break down our estimates into different groups based on the birth weights of the smaller twin. As can be seen, the estimated effect of a marginal increase in birth weight is quite similar for pairs with at least one low birth weight (<2500g) twin and those with only normal birth weight (≥2500g) twins; the estimate for the former is 0.428, and for the latter it is 0.526, and the two pooled coefficients are not

[36] 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 online appendix figure A8.

[37] 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: 0.943).

[38] At very small discordances of less than three or four percent, the estimates are too noisy to obtain a meaningful result. We exclude the very small discordances, therefore, so that the results for more meaningful discordances are more straightforward to present and observe.

statistically distinguishable from one another. Likewise, the estimated effects reported in rows 4 and 5 of the table for twin pairs with at least one very low birth weight (<1500g) and those where the smallest twin is low birth weight (1500-2499g) twins do not vary substantially across these groups. The estimated effects for very low birth weight, low birth weight and normal weight are, respectively, 0.432, 0.431 and 0.526. The table also presents other specifications, such as birth weight measured linearly, and birth weight interacted with the population demeaned mean birth weight in the twin pair, and all sets of results paint the same fundamental picture. [39]

VI. 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 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 the twin pair attends a high quality

[39] Additional formal tests supporting the linear in log birth weight specification are described in Figlio et al. (2013).

school instead of a low quality school? In other words, does school quality remediate the effect of early health deficits?

To answer this question, we measure school quality in six different ways. All are based on test scores; however, the available evidence (e.g., Chetty et al., 2011, Chetty, Friedman, and Rockoff 2013) suggests that measures of school or teacher quality based on test scores correlate strongly with later-life outcomes. 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. We stratify schools based on average proficiency levels and average student gains from year to year. [40] In addition, because jurisdictions have made very different determinations about what it means to be a "good" school, we have coded, to the closest degree possible in our data, three other highly-publicized state/city school grading systems that weight measures of school quality in substantially different ways – the systems in Indiana, Louisiana, and New York City.

[Insert Table 4 Here]

[Insert Table 5 Here]

The results of the school quality analyses are presented in tables 4 and 5 (similarly to table 3 we present mean group test scores and birth weight in the print appendix table A1). The first panel of table 4 shows estimates separately for

[40] 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 0.71 and with growth in achievement at 0.23, while the average achievement correlates with achievement growth at 0.03.

twins who attended schools that received an A, a B, and a C or below. 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 and singletons who attend higher-rated schools tend to have heavier birth weights than those attending lower-rated schools. But while there are relationships between school grade, birth weights, and test scores, there is no monotonic relationship in the association between birth weight and test scores: The estimated effect of birth weight is largest among twins who attend schools receiving a grade of B (0.499). 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 (0.458). These coefficients are not statistically distinguishable from one another. The point estimates are even closer together for singletons, where the estimated coefficient on birth weight varies between 0.273 and 0.284 (0.224 to 0.237 for sibling pairs) and the estimated coefficient on birth weight conditional on gestational length ranges from 0.377 to 0.413 (0.276 to 0.333 for siblings).

Florida's school grades are based in large measure on the school's average FCAT scores and the school's average student-level FCAT score improvements. The second and third panels of table 4 explicitly subdivide schools based on these dimensions. We find that regardless of whether schools are stratified by average levels of FCAT scores or average score improvements, the estimated effects of birth weight are present and approximately the same. For instance, the estimated marginal effect of log birth weight for twins attending schools with above-median FCAT scores is 0.426, versus 0.437 for twins attending schools with below-median FCAT scores, and the estimated marginal effect twins attending a school that had above-median year-to-year gains in FCAT scores is 0.427, versus 0.453 for schools with below-median gains in FCAT scores.

Applying other jurisdictions' school grading formulas to Florida's data, as reported in table 5, does not change the fundamental conclusion regarding school

quality. We break the Florida school rankings based on each of the three state alternative grading systems into thirds and find several consistent patterns: First, the estimated relationship between log birth weight and student test scores is strong and present in all cases. Second, there is rarely a monotonic relationship observed between the measure of school quality and the coefficient on log birth weight, whether it is derived from a twin fixed effect model, a sibling fixed effect model or from a singletons model controlling for gestational length or from a singletons model without controlling for gestation. Third, in the rare cases in which there exists a monotonic relationship, in one case (singletons in New York City) the pattern runs counter to that of the other two (sibling fixed effects in Indiana and Louisiana), and in all cases the coefficient estimates are very similar.[41]

Given that we observe larger estimated effects of birth weight for higher socio-economic status families than for lower socio-economic status families, and since higher socio-economic status families tend to select into higher-rated schools, it is possible that our finding of no relationship between measured school quality and the estimated effect of birth weight is biased due to these differentials. To investigate this possibility, we repeat the school grades analysis but further stratify the estimated effects of birth weight by predicted socio-economic status using the same approach that we followed to generate figure 8. These results are presented in online appendix table A3. We continue to observe strong, positive relationships between log birth weight and test scores for all school grade levels and all predicted socio-economic groups. In addition, there continues to be no consistent pattern in these estimated relationships across school grades. For the twin fixed effect model, the smallest estimated effects are seen in A schools in two of the three socio-economic groups (the lower and middle SES groups), but

[41] The relationship between gestational length and test scores is monotonic in measured school quality, but the results across measured school quality are always similarly-sized, consistent with our overall findings.

the patterns are different for singletons. It appears, therefore, that the differential selection of higher-SES families into higher-rated schools is not responsible in a substantial way for our finding that school quality appears to not substantively affect the relationship between birth weight and student outcomes.[42]

In summary, the evidence appears to indicate that the effect of birth weight on test scores does not vary substantially 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, 1992a, Card & Krueger, 1992b, Krueger & Whitmore, 2001, Chetty et al., 2011, Chetty, Friedman, and Rockoff 2013) of schooling have been shown to have significant positive impacts on earnings and other outcomes. If attending a better school improves all students' outcomes in parallel but does not completely remediate the effects of early health deficits on cognitive development, it may be that schools currently lack the resources or information necessary to fully remediate these deficits.

An alternative view of the results is 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 between twins or siblings is probably far more noticeable to parents than to classroom teachers. To parents a 15 percent difference in twins' or siblings' birth weight would be noticeable, but to a teacher nine to fourteen years later children's initial birth weights would be insignificant compared to the cognitive achievement

[42] We have also estimated models in which we control for log birth weight interacted with observable maternal and socio-economic characteristics. Our results regarding no apparent relationship between school quality measures and the estimated effect of log birth weight are fundamentally unchanged when we further condition on these interaction terms.

she observes in the classroom. Even differences in cognitive achievement resulting from large discordances in birth weight among twins or siblings probably appear to the teacher to be the result of temperamental differences. 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 or siblings – or, for that matter, similar children under a different dimension – similarly. The lack of relative improvement of children with poor neonatal health in better-rated schools may not indicate that it is impossible to remediate. Rather, it may indicate that it is not done, or at least not done systematically.

VII. 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 highly developed context, provide the first comprehensive study of the differential effects on a wide range of 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: Children with higher birth weight enter school with a cognitive advantage that appears to remain stable through the elementary and middle school years. The birth weight-related patterns in test score performance observed in twins are also seen in the overall population of singletons. 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 other background characteristics. The estimated effects of neonatal health are of roughly the same magnitude throughout the tested

grades as they are at the beginning of kindergarten (Figlio et al., 2013), and even as they are in very early childhood (Hart, 2008).[43] 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.

This pattern persists despite parental attempts to provide different experiences to their different children in early childhood. Bharadwaj, Eberhard, and Neilson (2013) and Hsin (2012), for example, find evidence that parents tend to invest more in lower birth weight children than they do in higher birth weight children, indicating a desire for remediation. While our administrative data do not offer the types of survey data used in those two papers, we see evidence of parents actively and simultaneously making different choices for their *twins*, suggesting that parents recognize developmental differences in their children and seek to remediate these differences in early childhood. It is reasonably common in Florida for parents to send one twin to preschool but not the other (true in 7.6 percent of twin pairs and 8.9 percent of twin pairs in which the birth weight discordance is greater than 20 percent). In 9.2 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 (1.2 percent of twin pairs with

[43] Hart's (2008) study of a much smaller set of twins in the ECLS-B finds estimated effects of birth weight on the Bayley Scales of Infant Development that are close in effect size to those presented in our paper.

discordance greater than 20 percent) parents "redshirt" one twin but not the other – starting twins in school at different ages.[44]

Children with poor neonatal health who come from highly-educated families perform much better than those with good neonatal health who come from poorly-educated families, indicating that "nurture" can at least partially overcome "nature." Indeed, this finding is very much in keeping with the literature on the positive relationship between household income and health status in childhood and adulthood (see, e.g., Case, Lubotsky and Paxson, 2002). Still, the fact that these initial biological factors are not fully overcome for even the most affluent and educated families – and, indeed, that the estimated effects of log birth weight are actually somewhat higher for these families – is consistent with the notion that parental inputs and neonatal health are complements rather than substitutes. While what exactly parents do to successfully remediate initial biological disadvantage and what schools and parents could potentially do in early childhood and the early elementary grades and beyond to continue to remediate are open questions, this study provides numerous indications that poor neonatal health establishes a stable trajectory for children's cognitive development.

These findings have potential implications for both health and education policy and practice. While it is premature to suggest specific policy responses based on this work, these findings indicate some potentially fruitful places to look for additional evidence. On the health side, for example, it will be valuable to learn whether improvements in earnings by families with pregnant women, improved maternal nutrition, or reduced maternal stress – all factors associated with higher birth weight – also translate to better cognitive outcomes in childhood. On the education side, it will be important to learn whether the relationship between birth weight and cognitive outcomes is attenuated in cases in which health and

[44] In cases of differential redshirting, parents are slightly more likely to redshirt the lighter twin than they are to redshirt the heavier twin. We discuss differential redshirting in greater detail in Figlio et al. (2013).

education providers have more interaction, such as in the case of children who participate in early intervention pre-kindergarten programs. Understanding these types of relationships will help us to modify the mechanisms through which neonatal health affects cognitive outcomes in childhood and adulthood, and guide health and education policy and practice.

APPENDIX

[Insert Table A1 Here]

REFERENCES

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

Antsaklis, Aris, Fotodotis M. Malamas, and Michael Sindos. 2013. "Trends in twin pregnancies and mode of delivery during the last 30 years: inconsistency between guidelines and clinical practice", *Journal of Perinatal Medicine* 41(4):355-64.

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. 2013. "Health at Birth, Parental Investments and Academic Outcomes" Unpublished.

Bitler, Marianne. 2008. "Effects of Increased Access to Infertility Treatment on Infant and Child Health: Evidence from Health Insurance Mandates" Unpublished.

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.

Blickstein, Isaac, and Robin Kalish. 2003. "Birthweight Discordance in Multiple Pregnancy", *Twin Research* 6: 526-531.

Breathnach, Fionnuala, and Fergal Malone. 2012. "Fetal Growth Disorders in Twin Gestations", *Seminars in Perinatology* 36:171-181.

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. 1992a. "Does School Quallity Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*. 100(1): 1-40.

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

Case, Anne, Darren Lubotsky, and Christina Paxson. 2002. "Economic Status and Health in Childhood: The Origins of the Gradient", *American Economic Review* 92(5): 1308-1334.

Chay, Kenneth, Jonathan Guryan, and Bhashkr Mazumder. 2009. "Birth Cohort and the Black-White Achievement Gap: The Roles of Access and Health Soon After Birth", National Bureau of Economic Research Working Paper #15078.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "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. 2013. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood", National Bureau of Economic Research Working Paper # 19424.

Conley, Dalton, Kate Strully, and Neil G. Bennett. 2003. "A Pound of Flesh or Just Proxy? Using Twin Differences to Estimate the Effect of Birth Weight on Life Chances", National Bureau of Economic Research Working Paper # 9901.

Conti, Gabriella, and James Heckman. 2010. "Understanding the Early Origins of the Education–Health Gradient: A Framework That Can Also Be Applied to Analyze Gene–Environment Interactions", *Perspectives in Psychological Science* 5: 585–605.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation,"

American Economic Review 97 (2): 31-47.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation," in Eric A. Hanushek and Finis Welch (eds.) *Handbook of the Economics of Education* Vol. 1: 697-812.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth. 2013. "The Effects of Poor Neonatal Health on Children's Cognitive Development", National Bureau of Economic Research Working Paper # 18846.

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.

Guryan, Jonathan, Erik Hurst, and Melissa Kearney. 2008. "Parental Education and Parental Time Use," *Journal of Economic Perspectives* 22(3): 23-46.

Hart, Cassandra. 2008. "Parenting and Child Cognitive and Socioemotional Development: A Longitudinal Twin Differences Study" Unpublished.

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.

Hoynes, Hilary W., Douglas L. Miller, and David Simon. 2014. "Income, the Earned Income Tax Credit, and Infant Health", *American Economic Journal: Economic Policy*, forthcoming.

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

Johnson, Rucker C., and Robert F. Schoeni. 2011. "The Influence of Early-Life Events on Human Capital, Health Status, and Labor Market Outcomes Over the Life Course", *Advances in Economic Analysis and Policy* 11(3): 1-55.

Kent, Etaoin M., et al. 2011. "Placental Cord Insertion and Birthweight Discordance in Twin Pregnancies: Results of the National Prospective ESPRiT Study", *American Journal of Obstetrics and Gynecology* 205: 376.e1-7.

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.

Ladd, Helen, Clara Muschkin, and Kenneth Dodge. 2012. "From Birth to School: Early Childhood Initiatives and Third Grade Outcomes in North Carolina" Unpublished.

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.

Lau, Carissa, Namasivavam Ambalavanan, Hrishikesh Chakraborty, Martha S. Wingate, and Waldemar A. Carlo. 2013. "Extremely Low Birth Weight and Infant Mortality Rates in the United States", *Pediatrics* 131: 855-60.

Luu, Thuy M., and Betty 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. 2014. "Economic Growth, Comparative Advantage, and Gender Differences in Schooling Outcomes: Evidence from the Birthweight Differences of Chinese Twins", *Journal of Development Economics*, forthcoming

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.

FIGURES

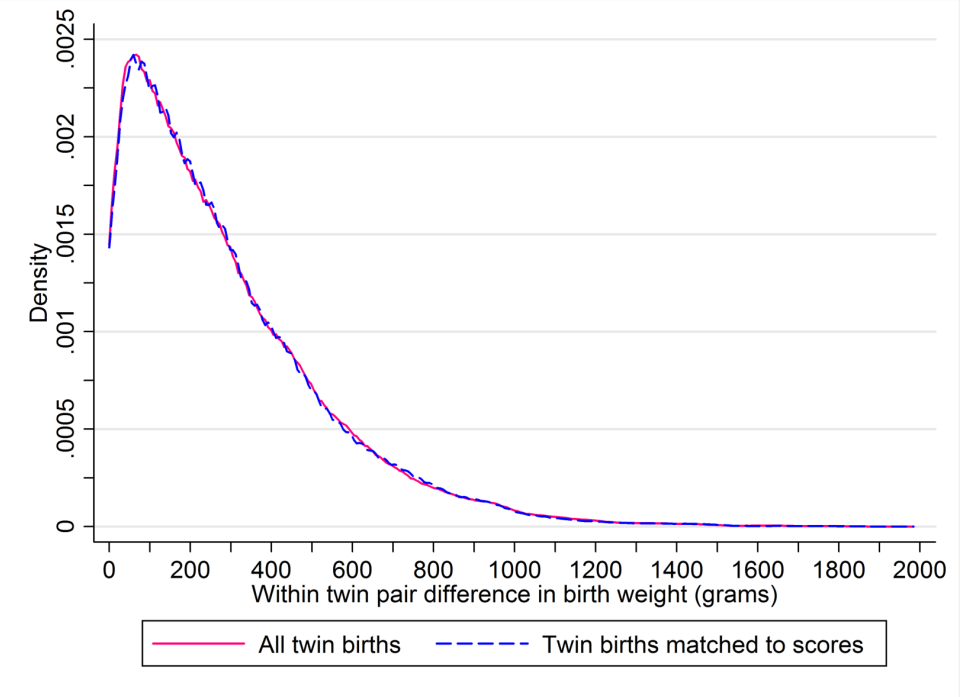

FIGURE 1. DISCORDANCE IN BIRTH WEIGHT BETWEEN TWINS BORN IN FLORIDA BETWEEN 1992 AND 2002

Notes: 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.

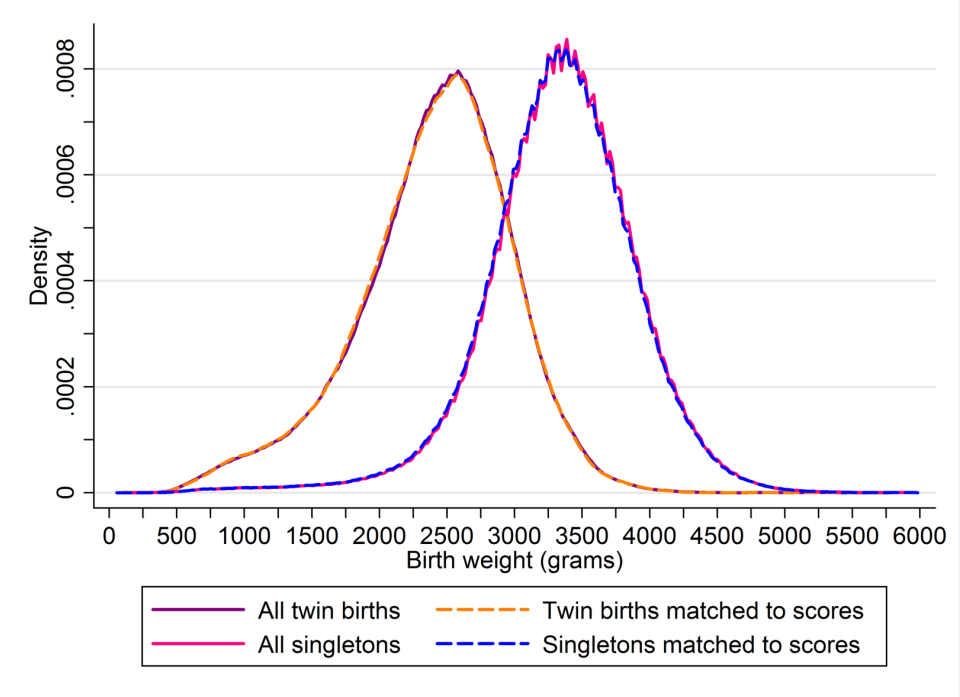

FIGURE 2. DIFFERENCE IN BIRTH WEIGHT DISTRIBUTIONS BETWEEN SINGLETONS AND TWINS BORN IN FLORIDA BETWEEN 1992 AND 2002

Notes: 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.

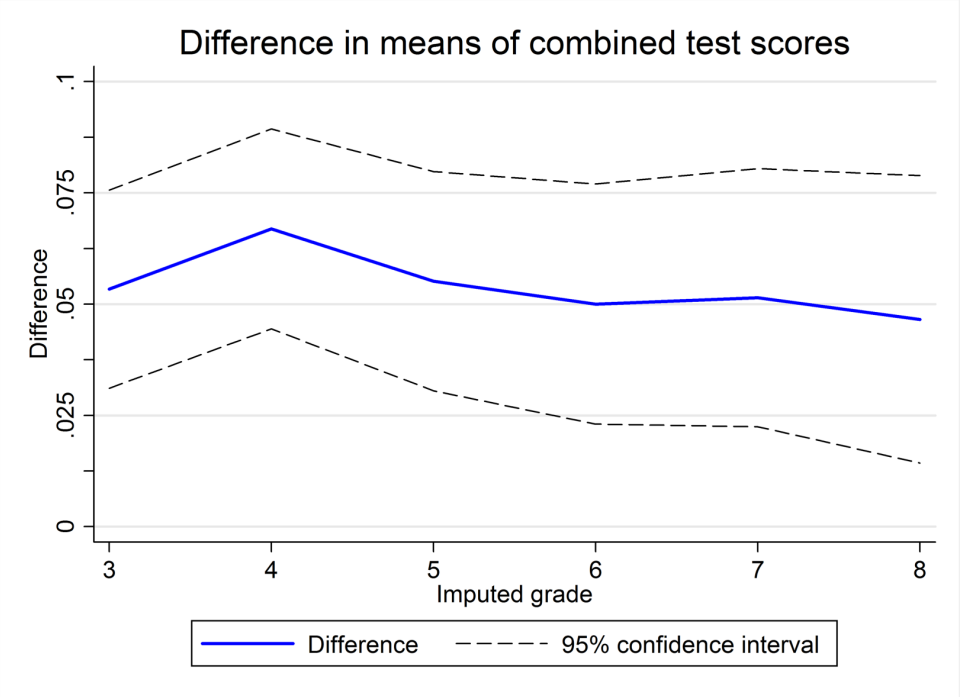

FIGURE 3. AVERAGE WITHIN-TWIN-PAIR DIFFERENCE IN TEST SCORES BETWEEN HEAVIER AND LIGHTER TWINS

Notes: 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.

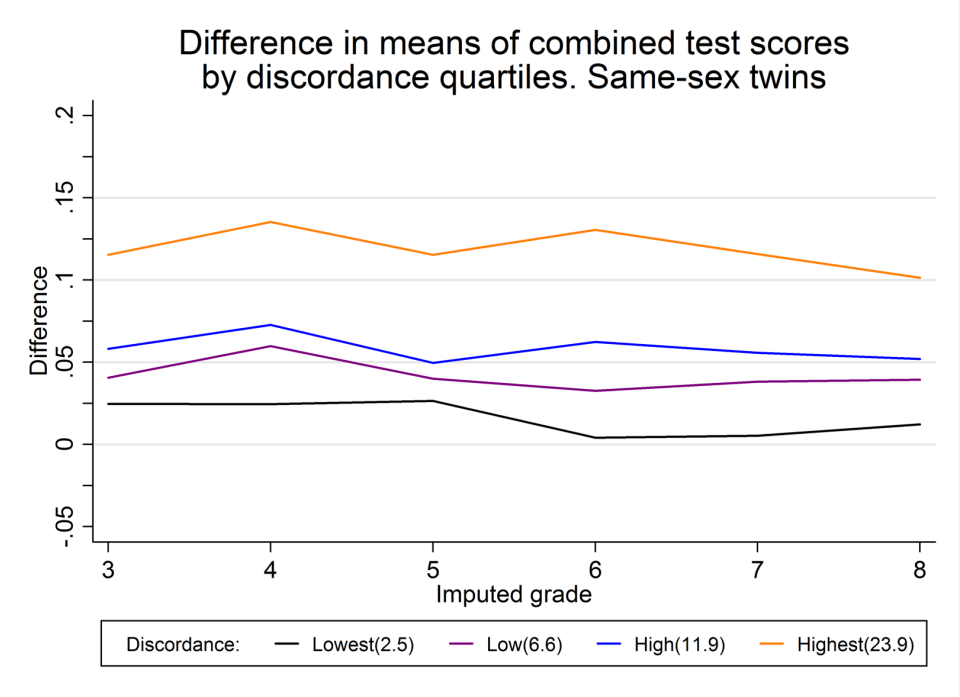

FIGURE 4. MEANS OF SCORES BY DISCORDANCE QUARTILES

Notes: Figure 4 plots difference between the mean test score of heavier and lighter twin from each pair in each grade for four quartiles of discordance in birth weight. 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. Discordance is calculated as the difference between heavier and lighter twin birth weight over the weight of the heavier twin. Mean discordance for each group in parentheses.

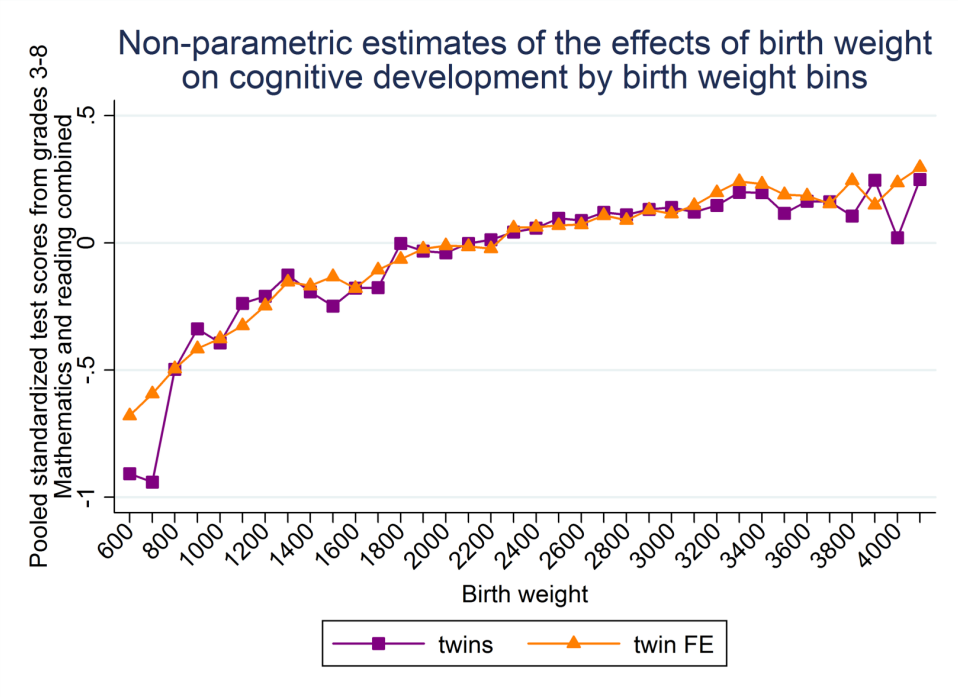

FIGURE 5. NON-PARAMETRIC RELATIONSHIP BETWEEN BIRTH WEIGHT AND TEST SCORES

Notes: Figure 5 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 of 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.

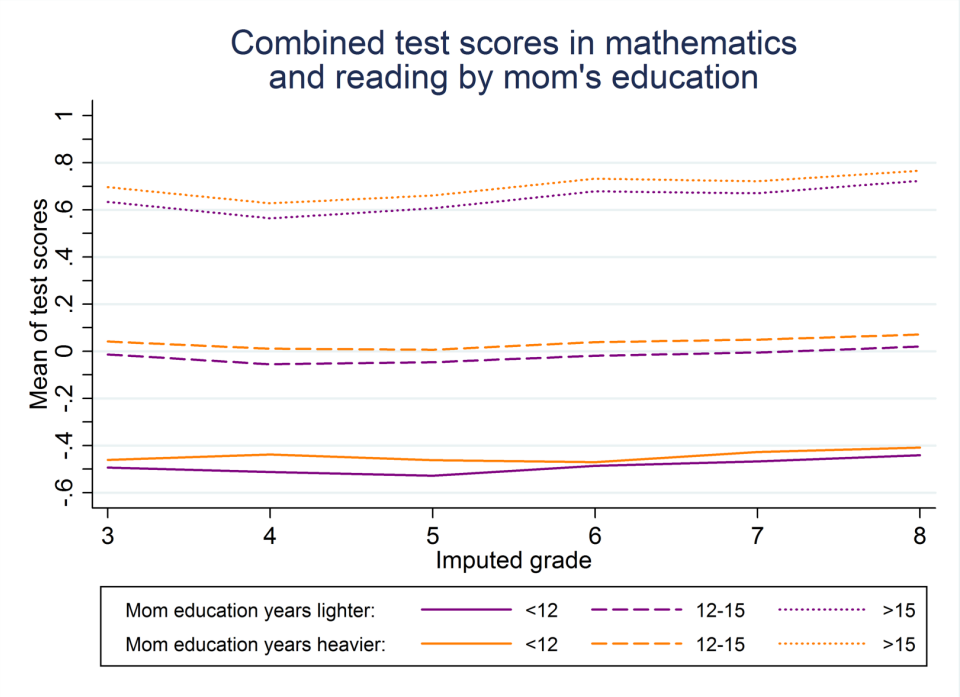

FIGURE 6. AVERAGE WITHIN TWIN PAIR DIFFERENCE IN TEST SCORES BETWEEN THE HIGHER BIRTH WEIGHT AND THE LOWER BIRTH WEIGHT TWIN BY MATERNAL EDUCATION CATEGORIES

Notes: Figure 6 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.

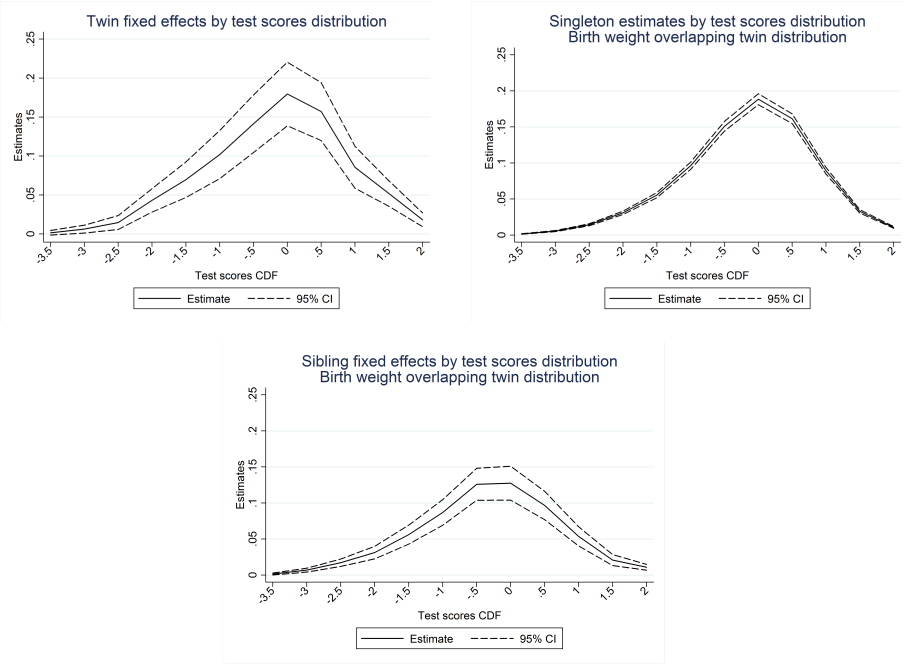

FIGURE 7. ESTIMATED EFFECTS OF BIRTH WEIGHT ON THE POSITION IN THE TEST SCORE DISTRIBUTION

Notes: Figure 7 plots estimated effects of log birth weight on the CDF of test scores. Specifically, the top-left panel plots coefficients on log birth weight from a series of standard twin FE regressions in which the dependent variables are indicators marking various points in the CDF of test scores (e.g. greater than -3.5, greater than -3, etc.). The top-right panel plots estimates from analogous regressions that include singletons with birth weights that overlap with the twin birth weight distribution. The bottom-center panel plots estimates from analogous sibling fixed effects regressions conditional on gestation that include singletons with birth weight that overlap with the twin birth weight distribution.

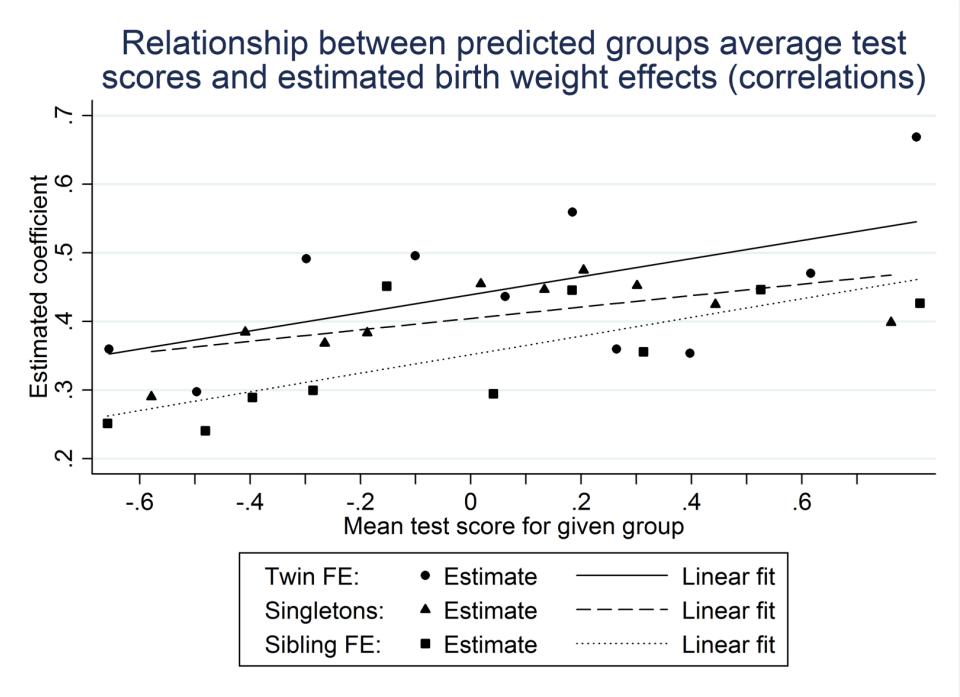

FIGURE 8. AVERAGE TEST SCORES AMONG GROUPS AND ESTIMATED BIRTH WEIGHT EFFECTS

Notes: Figure 8 plots the estimates for the 10 predicted groups based on the regression of test scores on maternal race, ethnicity, immigrant origin, marital status, education, age categories and income indicators. These groups are not overlapping. In this graph income from 1992 and 1993 is imputed based on observables. Groups are calculated only for individuals with all information available and for all singletons and siblings with birth weight in a range of 847 to 3600 grams.

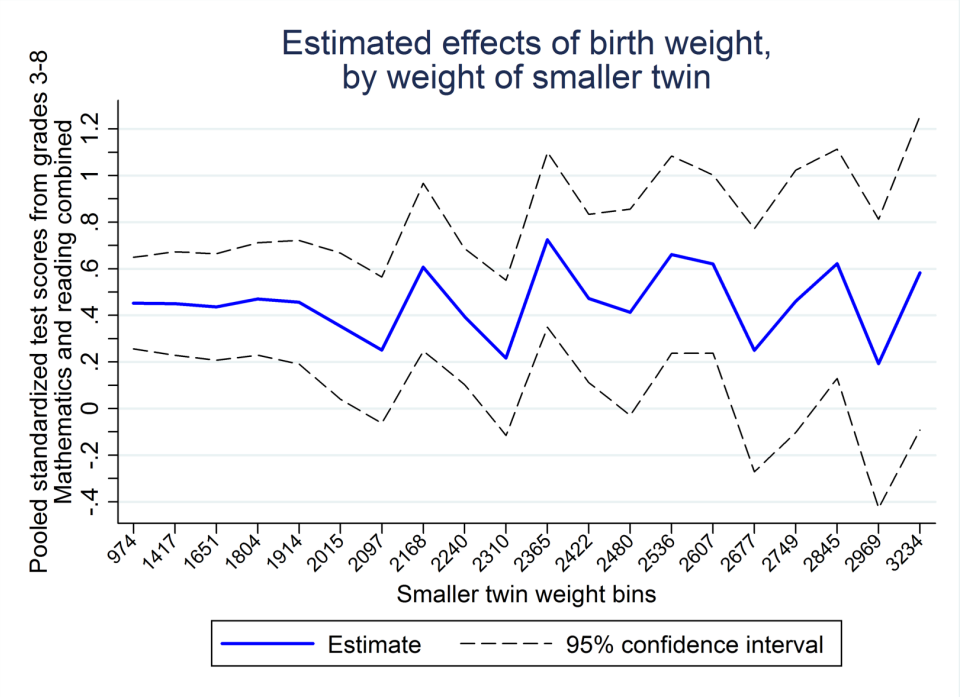

FIGURE 9. ESTIMATED EFFECTS OF BIRTH WEIGHT, BY WEIGHT OF SMALLER TWIN

Notes: Figure 9 plots coefficient estimates from a twin FE regression where the dependent variable is the mean test score and the independent variables are the products 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 smaller twin birth weight in each of the 20 bins.

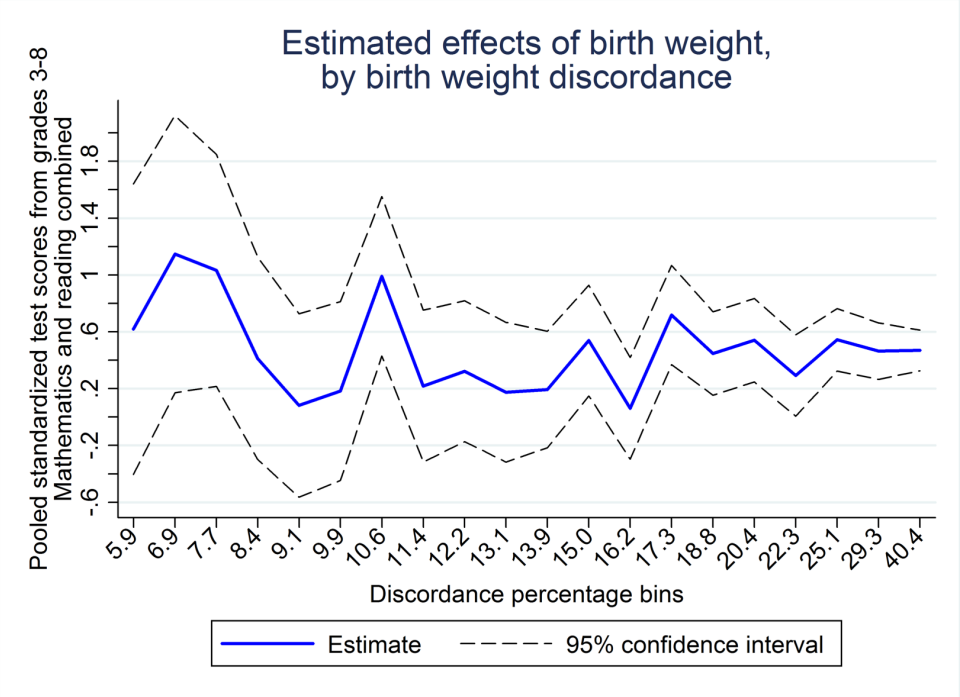

FIGURE 10. ESTIMATED EFFECTS OF BIRTH WEIGHT, BY BIRTH WEIGHT DISCORDANCE

Notes: Figure 10 plots coefficient estimates from a twin FE regression where the dependent variable is the mean test score and the independent variables are the products of log birth weight with indicators for 20 bins reflecting birth weight discordance 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 twin pair percentage discordance.

TABLES

TABLE 1—REPRESENTATIVENESS OF THE FLORIDA TEST SCORE AND 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	23.9	18.0
High school dropout	20.9	22.5	23.3	15.5
High school graduate	58.6	60.0	60.5	61.5
College graduate	20.1	17.1	15.8	23.1
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	26.1	24.8	24.4	31.8
Age 36 or above	9.8	9.3	9.2	13.6
Foreign-born	23.4	22.9	23.2	18.0
Married at time of birth	64.8	62.2	60.9	68.4
Number of children	2,047,663	1,652,333	1,334,006	28,434

Notes: 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 fulfilling. We restrict columns (3) and (4) only to observations that include full information on birth certificate.

TABLE 2—ESTIMATED EFFECTS OF BIRTH WEIGHT ON COGNITIVE DEVELOPMENT

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pooled		Imputed grade					
	OLS	FE	3	4	5	6	7	8
Twins *(Average of mathematics and reading)*: Estimates on ln(birth weight)								
All twins	0.285***	0.443***	0.444***	0.526***	0.431***	0.428***	0.390***	0.376***
	(0.022)	(0.039)	(0.043)	(0.045)	(0.047)	(0.053)	(0.057)	(0.061)
	[126,636]		[28,434]	[26,508]	[22,970]	[19,340]	[16,186]	[13,198]
Same sex twins	0.300***	0.452***	0.463***	0.532***	0.411***	0.469***	0.402***	0.368***
	(0.027)	(0.043)	(0.050)	(0.053)	(0.053)	(0.059)	(0.062)	(0.066)
Opposite sex twins	0.259***	0.421***	0.399***	0.513***	0.475***	0.330***	0.360***	0.390***
	(0.038)	(0.082)	(0.086)	(0.088)	(0.097)	(0.112)	(0.122)	(0.136)
Singletons (*Average of mathematics and reading)*: Estimates on ln(birth weight) and gestation								
Ln(birth weight)	0.285***	-	0.305***	0.289***	0.292***	0.281***	0.262***	0.261***
	(0.004)		(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
	[5,752,665]		[1,254,821]	[1,181,590]	[1,040,814]	[888,895]	[756,478]	[630,067]
Ln(birth weight) | gestation weeks	0.332***	-	0.345***	0.336***	0.337***	0.328***	0.313***	0.316***
	(0.005)		(0.005)	(0.005)	(0.006)	(0.006)	(0.007)	(0.007)
Ln(birth weight) | gestation weeks [overlapping]	0.421***	-	0.430***	0.424***	0.428***	0.421***	0.399***	0.406***
	(0.007)		(0.008)	(0.008)	(0.009)	(0.009)	(0.010)	(0.011)
Gestation weeks	0.013***	-	0.015***	0.013***	0.013***	0.012***	0.011***	0.010***
	(0.000)		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Siblings (*Average of mathematics and reading)*: Estimates on ln(birth weight) and gestation								
Ln(birth weight)	0.277***	0.238***	0.263***	0.254***	0.241***	0.219***	0.179***	0.178***
	(0.009)	(0.011)	(0.012)	(0.013)	(0.015)	(0.017)	(0.021)	(0.026)
	[1,110,206]		[294,782]	[267,751]	[212,294]	[156,910]	[109,883]	[68,586]
Ln(birth weight) | gestation weeks [overlapping]	0.403***	0.317***	0.345***	0.335***	0.315***	0.344***	0.227***	0.200***
	(0.018)	(0.022)	(0.024)	(0.025)	(0.028)	(0.033)	(0.039)	(0.050)
Gestation weeks	0.012***	0.008***	0.009***	0.009***	0.008***	0.005***	0.006***	0.005**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)

Notes: Columns (1) and (2) present pooled grade three through eight results for OLS, twin and sibling-FE models. Columns (3) to (8) present OLS, twin and sibling-FE estimates separately for each of the 6 grades. Each coefficient comes from a separate regression. Sample sizes in square brackets reflect number of individual observations in each regression; only twin pairs where both twins are observed with test scores in each grade are included; only siblings where at least two siblings are observed with test scores in each grade are included. All singletons are included except for the second to last estimate for singletons where only singletons with birth weight in range 847 to 3600 grams. Siblings could be identified only in about half of the population. We include all siblings that have test scores in given grade. In second to last column we focus only on siblings where the birth weight ranges from 847 to 3600 grams. This restriction provides overlapping distribution of birth weight among twins and singletons. The dependent variables are averaged 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 and within-twin pair birth order. OLS estimates further controls for infant birth month and year, marital and immigration status, race and ethnicity, indicators for maternal age (each for one year), education (high school dropout, high school graduate, college graduate) and number of births (each for one birth). Sibling fixed effects estimates further control for birth order within a family. Naturally time invariant characteristics of the mothers are dropped in sibling fixed effects specifications. In siblings regressions we additionally control for birth order within the sibling pair observed in our data. Standard errors in all twin estimates are clustered at twin pair level. Standard errors in singleton estimates are clustered at individual level in pooled regressions (column (1)) while heteroskedasticity robust standard errors are calculated in columns (3) to (8) where there is just one observation per individual. Standard errors in all sibling estimates are clustered at mother level.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

TABLE 3—ESTIMATED EFFECTS OF BIRTH WEIGHT ON COGNITIVE DEVELOPMENT BY CHILD AND MOTHER CHARACTERISTICS

Sample	(1) Twins	(2)	(3) Singletons	(4)	(5)	(6) Siblings	(7)
	Birth weight	Birth weight	Birth weight | gestation	Gestation	Birth weight	Birth weight | gestation	Gestation
Boys	0.454***	0.296***	0.440***	0.013***	0.230***	0.321***	0.007***
	(0.068)	(0.005)	(0.011)	(0.001)	(0.022)	(0.044)	(0.002)
Girls	0.449***	0.276***	0.407***	0.013***	0.223***	0.291***	0.008***
	(0.052)	(0.005)	(0.010)	(0.000)	(0.021)	(0.037)	(0.002)
No medical	0.449***	0.296***	0.437***	0.011***	0.249***	0.331***	0.007***
problems	(0.048)	(0.005)	(0.009)	(0.000)	(0.015)	(0.028)	(0.001)
Medical	0.422***	0.249***	0.372***	0.015***	0.244***	0.319***	0.011***
problems	(0.066)	(0.006)	(0.013)	(0.001)	(0.032)	(0.063)	(0.003)
White	0.464***	0.293***	0.457***	0.011***	0.244***	0.346***	0.006***
	(0.045)	(0.005)	(0.009)	(0.000)	(0.015)	(0.030)	(0.001)
Black	0.392***	0.262***	0.344***	0.015***	0.232***	0.282***	0.011***
	(0.082)	(0.006)	(0.013)	(0.001)	(0.017)	(0.033)	(0.002)
Non-Hispanic	0.436***	0.283***	0.426***	0.012***	0.228***	0.304***	0.007***
	(0.044)	(0.004)	(0.008)	(0.000)	(0.013)	(0.025)	(0.001)
Hispanic	0.480***	0.270***	0.384***	0.012***	0.270***	0.357***	0.012***
	(0.079)	(0.008)	(0.015)	(0.001)	(0.023)	(0.046)	(0.002)
Non-immigrant	0.441***	0.284***	0.422***	0.012***	0.223***	0.292***	0.006***
	(0.044)	(0.004)	(0.008)	(0.000)	(0.013)	(0.024)	(0.001)
Immigrant	0.456***	0.255***	0.379***	0.013***	0.291***	0.411***	0.012***
	(0.077)	(0.008)	(0.015)	(0.001)	(0.024)	(0.048)	(0.002)
Education	0.358***	0.265***	0.368***	0.012***	0.229***	0.303***	0.008***
Below 12 yrs	(0.094)	(0.008)	(0.014)	(0.001)	(0.026)	(0.046)	(0.002)
12-15 yrs	0.439***	0.291***	0.436***	0.013***	0.225***	0.306***	0.008***
	(0.050)	(0.005)	(0.009)	(0.000)	(0.016)	(0.030)	(0.001)
Above 15 yrs	0.523***	0.256***	0.380***	0.013***	0.238***	0.418***	0.001
	(0.079)	(0.010)	(0.020)	(0.001)	(0.031)	(0.059)	(0.003)
Bottom	0.388***	0.289***	0.407***	0.015***	0.250***	0.287***	0.011***
	(0.076)	(0.007)	(0.013)	(0.001)	(0.020)	(0.038)	(0.002)
Middle	0.445***	0.269***	0.407***	0.012***	0.221***	0.339***	0.007***
	(0.072)	(0.007)	(0.014)	(0.001)	(0.024)	(0.047)	(0.002)
Top	0.447***	0.264***	0.400***	0.011***	0.239***	0.401***	0.004*
	(0.078)	(0.008)	(0.016)	(0.001)	(0.026)	(0.049)	(0.002)
Non-married	0.372***	0.269***	0.384***	0.013***	0.235***	0.284***	0.009***
	(0.076)	(0.006)	(0.011)	(0.001)	(0.018)	(0.034)	(0.002)
Married	0.482***	0.292***	0.439***	0.012***	0.259***	0.366***	0.007***
	(0.044)	(0.005)	(0.010)	(0.000)	(0.017)	(0.032)	(0.001)
Age below 22	0.372***	0.268***	0.373***	0.011***	0.195***	0.305***	0.005**
	(0.115)	(0.007)	(0.014)	(0.001)	(0.025)	(0.046)	(0.002)
22-29	0.444***	0.274***	0.415***	0.011***	0.249***	0.317***	0.009***
	(0.059)	(0.006)	(0.012)	(0.001)	(0.022)	(0.042)	(0.002)
30-35	0.490***	0.294***	0.446***	0.014***	0.228***	0.329***	0.006**
	(0.069)	(0.007)	(0.015)	(0.001)	(0.034)	(0.066)	(0.003)
Above 35	0.410***	0.326***	0.490***	0.018***	0.269***	0.335***	0.016***
	(0.104)	(0.012)	(0.024)	(0.001)	(0.054)	(0.119)	(0.005)

Notes: Column (1) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (2) to (4) present estimates for singleton population. Column (2) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (3) presents the correlation between pooled grades three through eight test scores and birth weight conditional on gestation for the sample of singletons that overlap in birth weight with twin population, i.e. birth weight in range 847 to 3600 grams. Column (4) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. Columns (5) to (7) present estimates for sibling population. Twins fixed effects regressions control for child gender and birth order. All singleton models include the

following controls: gender, month and year of birth dummies, marital and immigrant status, race and ethnicity, dummies for maternal education (3 categories), age and number of births. Sibling models further control for birth order within a family. Standard errors in column (1) are clustered at twin-pair level, in columns (2) to (4) at individual level while in columns (5) to (7) at mother level. Sample sizes are: 126 636 individual-years observations in column (1), 5,752,665 individual-year observations in columns (2) and (4), 4,025,893 individual-year observations in column (3), 1,110,206 individual-year observation in columns (5) and (7), 648,486 individual-year observations in column (6). There are fewer observations in zip code income because we do not observe these for years 1992 and 1993. There are fewer observations in racial breakdown because we exclude other races than Black or White from this comparison. There are fewer observations in maternal marital history breakdown because we miss information for some mothers.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

TABLE 4—RESULTS BY SCHOOL QUALITY MEASURES

Sample	(1) Twins	(2)	(3) Singletons	(4)	(5)	(6) Siblings	(7)
	Birth weight	Birth weight	Birth weight | gestation	Gestation	Birth weight	Birth weight | gestation	Gestation
			Awarded grade				
A	0.407***	0.273***	0.412***	0.012***	0.233***	0.333***	0.005***
	(0.042)	(0.004)	(0.009)	(0.000)	(0.014)	(0.027)	(0.001)
B	0.499***	0.284***	0.413***	0.012***	0.224***	0.305***	0.007***
	(0.063)	(0.006)	(0.011)	(0.001)	(0.022)	(0.043)	(0.002)
C & D & F	0.458***	0.275***	0.377***	0.014***	0.237***	0.276***	0.010***
	(0.076)	(0.006)	(0.012)	(0.001)	(0.021)	(0.040)	(0.002)
			Average proficiency				
Below median	0.437***	0.281***	0.395***	0.014***	0.230***	0.293***	0.010***
	(0.061)	(0.005)	(0.010)	(0.000)	(0.016)	(0.030)	(0.001)
Above median	0.426***	0.267***	0.404***	0.011***	0.240***	0.348***	0.005***
	(0.043)	(0.004)	(0.009)	(0.000)	(0.015)	(0.029)	(0.001)
			Growth in proficiency				
Below median	0.453***	0.286***	0.428***	0.012***	0.245***	0.324***	0.008***
	(0.044)	(0.004)	(0.008)	(0.000)	(0.014)	(0.026)	(0.001)
Above median	0.427***	0.284***	0.413***	0.013***	0.229***	0.281***	0.008***
	(0.045)	(0.004)	(0.008)	(0.000)	(0.014)	(0.026)	(0.001)

Notes: Column (1) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (2) to (4) present estimates for singleton population. Column (2) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (3) presents the correlation between pooled grades three through eight test scores and birth weight conditional on gestation for the sample of singletons that overlap in birth weight with twin population, i.e. birth weight in range 847 to 3600 grams. Column (4) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. Columns (5) to (7) present estimates for sibling population. Twins fixed effects regressions control for child gender and birth order. All singleton models include the following controls: gender, month and year of birth dummies, marital and immigrant status, race and ethnicity, dummies for maternal education (3 categories), age and number of births. Sibling models further control for birth order within a family. Standard errors in column (1) are clustered at twin-pair level, in columns (2) to (4) at individual level while in columns (5) to (7) at mother level. In the case of awarded grades since not all schools are awarded grades every year our sample consist of 123,886 observations used in models in column (1), 5,650,536 observations used in models in column (2) and (4), 3,952,642 observations used in models in column (3), 1,084,620 observations used in models in columns (5) and (7), and 632,125 observations used in column (6). In the case of average proficiency we use 125,936 observations in models in column (1), 5,731,434 observations in models in columns (2) and (4), 4,011,368 observations in models in column (3), 1,106,452 observations used in models in columns (5) and (7), and 646,284 observations used in column (6). In the case of growth in proficiency we use 125,566 observations in models in column (1), 5,716,150 observations in models in columns (2) and (4), 4,000,486 observations in models in column (3), 1,102,938 observations used in models in columns (5) and (7), and 644,010 observations used in column (6). The discrepancy between the samples in table 3 and table 4 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).

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

TABLE 5—RESULTS BY SCHOOL QUALITY MEASURES: RUNNING FLORIDA DATA THROUGH OTHER STATE SCHOOL GRADING SYSTEMS

State	Quality group	(1) Twins Birth weight	(2) Birth weight	(3) Singletons Birth weight | gestation	(4) Gestation	(5) Birth weight	(6) Siblings Birth weight | gestation	(7) Gestation
(1) New York City	Top	0.389***	0.270***	0.405***	0.011***	0.195***	0.265***	0.004***
		(0.049)	(0.005)	(0.010)	(0.000)	(0.018)	(0.035)	(0.002)
	Middle	0.491***	0.275***	0.407***	0.012***	0.233***	0.318***	0.008***
		(0.051)	(0.005)	(0.009)	(0.000)	(0.018)	(0.033)	(0.002)
	Bottom	0.484***	0.294***	0.418***	0.014***	0.251***	0.293***	0.011***
		(0.062)	(0.005)	(0.011)	(0.001)	(0.020)	(0.038)	(0.002)
(2) Louisiana	Top	0.399***	0.263***	0.403***	0.011***	0.232***	0.353***	0.005***
		(0.048)	(0.005)	(0.010)	(0.000)	(0.018)	(0.034)	(0.002)
	Middle	0.480***	0.283***	0.409***	0.013***	0.241***	0.319***	0.008***
		(0.054)	(0.005)	(0.010)	(0.000)	(0.018)	(0.035)	(0.002)
	Bottom	0.450***	0.267***	0.360***	0.015***	0.218***	0.250***	0.010***
		(0.104)	(0.008)	(0.015)	(0.001)	(0.028)	(0.052)	(0.002)
(3) Indiana	Top	0.401***	0.260***	0.395***	0.011***	0.217***	0.330***	0.005***
		(0.047)	(0.005)	(0.010)	(0.000)	(0.017)	(0.034)	(0.001)
	Middle	0.522***	0.286***	0.415***	0.013***	0.236***	0.290***	0.008***
		(0.054)	(0.005)	(0.010)	(0.000)	(0.019)	(0.034)	(0.002)
	Bottom	0.434***	0.276***	0.384***	0.015***	0.243***	0.274***	0.010***
		(0.097)	(0.007)	(0.014)	(0.001)	(0.026)	(0.049)	(0.002)

Notes: Column (1) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (2) to (4) present estimates for singleton population. Column (2) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (3) presents the correlation between pooled grades three through eight test scores and birth weight conditional on gestation for the sample of singletons that overlap in birth weight with twin population, i.e. birth weight in range 847 to 3600 grams. Column (4) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. Columns (5) to (7) present estimates for sibling population. Twin fixed effects regressions control for child gender and birth order. All singleton models include the following controls: gender, month and year of birth dummies, marital and immigrant status, race and ethnicity, dummies for maternal education (3 categories), age and number of births. Sibling models further control for birth order within a family. Standard errors in column (1) are clustered at twin-pair level, in columns (2) to (4) at individual level while in columns (5) to (7) at mother level. In the case of awarded grades since not all schools are awarded grades every year and not every system was functioning through the same time period our samples differ. New York system simulation consist of 107794 observations used in models in column (1), 4972962 observations used in models in column (2) and (4), 3471424 observations used in models in column (3), 850751 observations used in models in columns (5) and (7) and 493281 observations used in models in column (6). Louisiana system simulation consist of 108926 observations used in models in column (1), 5027615 observations used in models in column (2) and (4), 3508071 observations used in models in column (3), 850751 observations used in models in columns (5) and (7) and 493281 observations used in models in column (6). Indiana system simulation consist of 107798 observations used in models in column (1), 4973114 observations used in models in column (2) and (4), 3471516 observations used in models in column (3), 850751 observations used in models in columns (5) and (7) and 493281 observations used in models in column (6).

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

TABLE A1—MEAN TEST SCORES AND BIRTH WEIGHT FOR GROUPS USED IN TABLES 3, 4, AND 5

Sample	(1) Mean test score [Mean birth weight] Twins	(2) Singletons	Sample	(3) Mean test score [Mean birth weight] Twins	(4) Singletons
Table 3			30-35	0.280	0.305
Boys	0.049	0.051		[2467]	[3390]
	[2473]	[3397]	Above 35	0.342	0.306
Girls	0.101	0.119		[2480]	[3353]
	[2369]	[3274]	**Table 4**		
No medical problems	0.076	0.098	A	0.278	0.276
	[2457]	[3359]		[2437]	[3365]
Medical problems	0.074	0.041	B	-0.093	-0.039
	[2356]	[3259]		[2410]	[3323]
White	0.258	0.228	C & D & F	-0.399	-0.310
	[2457]	[3393]		[2375]	[3266]
Black	-0.465	-0.362	Below median	-0.339	-0.248
	[2318]	[3180]		[2382]	[3281]
Non-Hispanic	0.099	0.110	Above median	0.298	0.298
	[2413]	[3333]		[2442]	[3371]
Hispanic	-0.034	0.001	Below median	0.046	0.058
	[2454]	[3346]		[2421]	[3337]
Non-immigrant	0.074	0.079	Above median	0.100	0.106
	[2414]	[3334]		[2422]	[3335]
Immigrant	0.082	0.106	**Table 5**		
	[2452]	[3344]	New York City		
Education below 12 yrs	-0.475	-0.339	Top	0.305	0.291
	[2339]	[3252]		[2440]	[3363]
12-15 yrs	0.005	0.095	Middle	0.053	0.073
	[2430]	[3348]		[2427]	[3336]
Above 15 yrs	0.663	0.677	Bottom	-0.180	-0.120
	[2451]	[3417]		[2395]	[3308]
Bottom	-0.216	-0.138	Louisiana		
	[2393]	[3285]	Top	0.375	0.371
Middle	0.121	0.085		[2448]	[3381]
	[2410]	[3337]	Middle	-0.090	-0.024
Top	0.437	0.381		[2414]	[3325]
	[2434]	[3382]	Bottom	-0.489	-0.377
Non-married	-0.359	-0.234		[2365]	[3250]
	[2336]	[3237]	Indiana		
Married	0.273	0.277	Top	0.359	0.349
	[2459]	[3396]		[2448]	[3376]
Age below 22	-0.394	-0.207	Middle	-0.068	-0.006
	[2268]	[3237]		[2412]	[3328]
22-29	-0.005	0.076	Bottom	-0.450	-0.352
	[2419]	[3357]		[2369]	[3259]

Notes: Descriptive statistics for each group reported in tables 3, 4, and 5. These present mean combined mathematics and reading test scores as well as mean birth weight for twins and singletons, respectively.