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

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

Revised: October 7, 2013

We are grateful to the Florida Department of Health and Florida Department of Education for providing us access to merged education and health data for the purposes of this project, and for the technical support in interpreting the key variables described herein. Figlio and Roth appreciate the financial support of the National Science Foundation, as well as the U.S. Department of Education, Smith Richardson Foundation and the Bill and Melinda Gates Foundation through the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Karbownik appreciates the support of a Hedelius Scholarship that permitted him to participate in this research while visiting Northwestern University. Seminar and conference participants at the Association for Education Finance and Policy, CESifo (Munich), Cornell University, Dartmouth College, Duke University, Emory University, International Workshop on Applied Economics of Education (Catanzaro, Italy), National Bureau of Economic Research, National Center for the Analysis of Longitudinal Data in Education Research, Northwestern University, Roosevelt University, Stanford University, Tulane University, University of California-Irvine, University of Florida, University of Maryland, University of Notre Dame, University of Oklahoma, University of Texas, University of Wisconsin-Madison, and Uppsala University provided helpful comments and feedback. Hilary Hoynes and three anonymous referees provided many helpful suggestions. All errors of omission and commission are our own.

I. Introduction

A large literature documents the effects of neonatal health (commonly proxied by birth weight) on a wide range of adult outcomes such as wages, disability, adult chronic conditions, and human capital accumulation. A series of studies, conducted in a variety of countries, including Canada, 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. For that matter, we know very little about whether these effects vary heterogeneously across different demographic or socio-economic groups, and since the existing literature, when it mentions effect heterogeneity at all, rarely presents subgroup-specific findings, it is impossible given the extant literature to know whether early neonatal health and parental inputs are complements or substitutes. As such, while we have strong evidence from twin comparison studies that poor initial health conveys a disadvantage in adulthood, we have little information about the potential roles for policy interventions in ameliorating this disadvantage during childhood.

[1] Examples of influential previous research include Behrman and Rosenzweig (2004) on schooling and wages, Almond et al. (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 et al. (2007) on neonatal outcomes, height, IQ, high school completion, employment, earnings and next generation birth weight, for Norway; Oreopoulos et al. (2008) on neonatal outcomes, health outcomes in adolescence, educational attainment and social assistance take up, for Canada; Rosenzweig and Zhang (2012) on educational attainment, wages and weight for height, for China; and Torche and Echevarria (2011) on fourth-grade mathematics test scores, for Chile. In a current working paper, Bharadwaj et al. (2013) study fourth-grade test scores and grades in school, also in Chile.

The reason for these gaps in the literature involves data availability. The datasets that previous researchers have used to study the effects of poor neonatal health on adult outcomes (e.g., Scandinavian registry data, or data matching a mother's birth certificate to her children's birth certificates) do not include information on schooling and human capital measures during key developmental years. And even the small number of studies that investigate the effects of birth conditions on test scores rather than adult outcomes (Bharadwaj et al., 2013; Torche and Echevarria, 2011; Rosenzweig and Zhang, 2009) are in developing contexts (e.g., 1990s-era Chile and China) that either lack the diversity necessary to explore heterogeneous effects of poor neonatal health on cognitive development in a manner generalizable to most OECD countries, or a sufficient level of affluence to study these effects in a highly developed context.[2][3] And while the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) of children born in the United States in 2001 oversamples twins, this data set is too recent to investigate outcomes in late elementary school or adolescence, too small to study heterogeneous effects of birth weight, and does not include cognitive outcomes that have high stakes for children.

Another gap in the adult-outcomes literature is that the subjects of that literature are rather old at present; they were necessarily born in the 1970s and earlier. Given the advances in modern neonatology, it is reasonable to believe that poor neonatal health in the 21st century may bear little resemblance to poor neonatal health fifty years ago.[4]

[2] China and Chile have both been rapidly developing over the past two decades, and Chile is today a squarely middle-income country with GDP per capita that is 59 percent of the OECD average when measured in purchasing power parity and 39 percent of the OECD average when denominated by exchange rate. That said, Chile is much wealthier now than in the 1990s when the extant studies' children were born. This is important context because there have been significant and continuing advances in medical technology over the past several decades that have reduced the lower end of viable birth weights, and possibly changed the life chances of babies of all birth weights. Since access to these technologies is disproportionate to wealthy countries, it is important to examine the effect of poor neonatal health in a setting with widespread access to understand the effects in that context. Indeed, there is reason to believe that fewer Chilean children were viable during their time periods than at a similar time in the United States: Torche and Echevarria (2011) show that the mean birth weight amongst Chilean twins in 1997-1999 was 2500 grams, 107 grams heavier than Florida's average in 1997-1999.

[3] Neither Chilean study can investigate the dynamics of test scores through childhood, as Chile's staggered implementation of the SIMCE exam means that they can study fourth grade test scores for some but not all cohorts. Bharadwaj et al. (2013) use teacher-assigned grades, rather than a broadly-comparable outcome measure, as their dependent variable that varies across age.

[4] As one example of the temporal differences in neonatology, whereas 50 years ago the threshold for infant viability was around 1500 grams, today the threshold for viability in developed countries is as low as 500 grams or even lower (Lau et al., 2013). As such, it is independently valuable to study the effects of birth weight using a more contemporary set of births than those used in the existing literature.

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.[5]

We make use of a major new data source that can fill these gaps in the literature. We match all births in Florida from 1992 through 2002 to subsequent schooling records for those remaining in the state to attend public school. Florida is an excellent place to study these questions because it is large (its population of around 17 million compares to Norway, Denmark, and Sweden combined) and heterogeneous (44.3 percent of mothers are racial or ethnic minorities, and 22.5 percent of mothers were born outside the United States). In addition, Florida is well-known for having some of the strongest education data systems in the United States; Florida, North Carolina, and Texas established the most advanced statewide student longitudinal data systems in the United States during the first half of the 2000s, and Florida has been testing children annually from third through tenth grade for over a decade. For several cohorts, Florida also implemented a universal kindergarten readiness assessment that allows us to explore the effects of birth weight on children's cognitive outcomes as early as age five. In addition to superb education data quality, Florida offers another major advantage when attempting to match birth and school records: Because children born in Florida are immediately assigned a social security number, and because social security numbers are collected upon school registration, Florida presents the opportunity for particularly effective matches between birth and school records. This allows Florida's health and education agencies the ability to nearly perfectly match births to school records. As we describe in the next section, our match rate is almost identical to what we would have expected based on American Community Survey data. With these new data, we follow over 1.3 million singletons and over 14,000 pairs of twins from birth through middle school to study the relationship between birth weight and cognitive development.

This paper makes several principal contributions to the literature. First, this paper represents the first analysis of population-level data in a highly developed context to study the effects of neonatal health on cognitive development. Our use of population-

[5] 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).

level data is important because it permits us the opportunity to estimate heterogeneous effects across a wide variety of demographic and socio-economic dimensions, in order to address both the stability of results across background and the degree to which parental inputs and early health are complements or substitutes.[6] Such complementarity could be driven by parents with more resources investing more or less 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.

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 theoretically (e.g. consider the role complementarities play in the models of human capital accumulation of Cuhna et al. (2006), Cunha and Heckman (2007), Conti and Heckman (2010)), 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.

In addition, we use these population-level data to estimate the relationship between birth weight and student outcomes for both twins and singletons. We show that with a richly specified model that holds constant gestation length and restricts attention to the range of birth weights that account for the vast majority of twin births, it is possible to produce what appears to be an unbiased estimate of the effect of birth weight on test scores using the population of singletons. This finding then allows us to show how the results we find for twins, which have strong internal validity, show very similar patterns across subgroups to results estimated from singletons, suggesting the findings have broader external validity than previously known.

[6] We are certainly not the first paper to conduct heterogeneity analyses. 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 et al.'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 et al. (2013) only split by high school and over versus middle school or lower education.

Most uniquely, 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 in the most natural place where public policy can play a role in promoting cognitive development amongst children of this age. We seek to understand the degree to which school quality can help to overcome disadvantages associated with poor neonatal health.

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

II. A new data source

A. Description of the data set and match diagnostics

We make use of matched data for all children born in Florida between 1992 and 2002 and educated in a Florida public school afterward. For the purposes of this study, Florida's education and health agencies matched children along three dimensions: first and last names, date of birth, and social security number. Rather than conducting probabilistic matching, the match was conducted such that a child would be considered matched so long as (1) there were no more than two instances of modest inconsistencies (e.g., a last name of "Johnson" in the birth record but "Johnsen" or "Johnosn" in the school record, or a social security number ending in "4363" in the birth record but ending in "4336" in the school record); and (2) there were no other children who could plausibly be matched using the same criteria. Common variables excluded from the match were used as checks of match quality. These checks confirm a very high and clean match rate:

In the overall match on the entire (not just twin) population, the sex recorded on birth records disagreed with the sex recorded in school records in about one-one thousandth of one percent of cases, suggesting that these differences are almost surely due to typos in the birth or school records.

Between 1992 and 2002, 2,047,663 births were recorded by the Florida Bureau of Vital Statistics, including 22,625 pairs of twins. Of these children, 1,652,333 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 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.[7] Recall that a child can only be matched in the Florida data if he or she (1) is born in Florida; (2) remains in the state of Florida until school age; (3) attends a Florida public school; and (4) is successfully matched between birth and school records using the protocol described above. Reasons (1) through (3) are "natural" reasons why we might lose children from the match. Our calculations from the American Community Survey indicate that, amongst the kindergarten-aged children found in that survey who were born in Florida, 80.9 percent were remaining in Florida at the time of kindergarten and were attending public school.[8] We therefore conclude that the match rate is extremely high, and that nearly all potentially matchable children have been matched in our data.

[7] 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 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).

[8] Indeed, this figure is an overstatement of the true expected match rate because the American Community Survey includes only children who are still living in the United States at the time of kindergarten. Given that some children born in Florida leave the country in their first five years because of emigration, because they were born to non-immigrant visitors to the country, or because they were born to undocumented immigrants who returned to their home countries, the true expected match is somewhat below 80.9 percent.

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

It is still the case that the set of Florida-born children attending Florida public schools differs fundamentally from the set of all Florida-born children. People who leave the state of Florida might differ from those who remain, and children attending public schools might differ fundamentally from those who attend private schools. It is therefore important to gauge how comparable the matched population is to the overall population of twins and of singletons born in Florida. Though such a comparison is separate from how the matched data differ from the population, it is important to note that twins differ from singletons in important ways. Twins have a mean lower gestational age and birth weight than singletons, and have older and more educated mothers, as well as mothers who are more likely to be married. We discuss issues of external validity in the conclusion.

Table 1 presents some evidence regarding the overall representativeness of our population of twins, along a number of dimensions: maternal race and ethnicity, maternal education, maternal age, maternal immigrant status, and parental marital status. There are four columns in the table: The first column reflects the total population of children born in Florida; the second 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. The comparison between the first and second columns makes clear the costs associated with carrying out this type of analysis in the United States, where children are lost for matching if they cross state lines between birth and school or if they attend private school. We observe that the set of matched children are more likely to be black (24.8 percent of matched children versus 22.6 percent of all children) and less likely to have married mothers (62.2 percent versus 64.8 percent of all children). The mothers of matched children are more likely to be less educated (17.5 percent college graduate versus 20.5 percent overall, and 22.5 percent high school dropout versus 20.9 percent overall) and are moderately younger (23.6 percent aged 21 or below versus 22.0 percent overall, and 9.3 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. As can be seen, 3rd-grade test-takers are still lower in terms of socio-economic status than are all children appearing in public school data. The fact that matched children are of somewhat lower socio-economic status, and that those with 3rd-grade scores are somewhat lower again, is unsurprising, given the well-documented relationship between family income (or parental education) and private school attendance.[9] However, our findings of estimated relationships between birth weight and test scores that are remarkably similar across very dissimilar groups reduces some of the potential concerns regarding external validity.

The comparison between the third and fourth columns of table 1 demonstrates the consequences of making use of twin comparisons, as is standard in the literature. As can be seen, mothers of twins are quite different from 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 16.2 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.)[10] 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 highly 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

[9] 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.

[10] 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.

the two newborns, and sometimes the difference is quite substantial. In Florida, the average discordance in birth weight is 284 grams (0.63 pounds), or 11.8 percent of the average twin's birth weight of 2410 grams.[11] Figure 1 presents the distribution of discordance for all twins, as well as all twins matched to test scores. As can be seen, the two distributions are virtually identical, so even though twins remaining in Florida and attending public schools have different maternal characteristics than do twins who leave Florida or attend private schools, the identifying variation does not differ at all. 51.4 percent of twin pairs have birth weight discordance over 200 grams, and 16.8 percent have birth weight discordance over 500 grams. 45.1 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.[12]

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. One can easily observe that for both twins and singletons the birth weight distribution of children observed in the test score data is identical to the distribution of all children born in Florida. 53.2 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.

Note that the average birth weight in our population is considerably smaller than those in previously-published studies using children from highly developed countries born a generation or two earlier, and, as mentioned above, those using children born contemporaneously in middle-income and developing countries. Consider the case of twins: The mean birth weight in Behrman and Rosenzweig's (2004) study of Minnesota

[11] 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.

[12] 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.

twins born 1936-1955 was 2557 grams; for Royer's (2009) California twins born 1960-1982, Black, Devereux, and Salvanes's (2007) Norway twins born 1967-1981, and Oreopoulos et al.'s (2009) Manitoba twins born 1978-1985, the mean birth weights were 2533 grams, 2598 grams, and 2517 grams, respectively. The lower mean birth weight in our sample is almost surely the result of improvements in medical technology that allow lower birth weight babies to survive longer. This change in technology and shift in the birth weight distribution highlight another benefit of studying recently-born children.[13]

III. 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. Following Almond, Chay and Lee (ACL, 2005) and Black, Devereaux and Salvanes (BDS, 2007), let

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

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 (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

[13] Note also that the rate of the twinning in the developed world increased 76 percent between 1980 and 2009, most likely owing to increased use of assisted reproduction methods (Ananth and Chauhan, 2012).

Florida.[14] For ease of presentation, we average standardized reading and mathematics FCAT scores for our dependent variable, but our results are presented separately for reading and mathematics, and the test-specific results are available on request. The regressor of interest, ln(*bw*), is the natural logarithm of birth weight in grams. In section 6 we present results from specifications other than the linear-in-log model, but the linear-in-log model appears to fit the data well.

Ordinary Least Squares (OLS) estimation of (1) would produce biased estimates of β if ϕ_{jk} were correlated with $\ln(bw)_{ijk}$ – in other words, if there were unobservable determinants of cognitive ability that were correlated with birth weight. To address the potential bias due to correlation between ϕ_{jk} and $\ln(bw)_{ijk}$, we estimate a twin fixed effect model. Twins necessarily share the same x_{jk} and ϕ_{jk}. A twin fixed effect model essentially differences out any mother- or birth-specific confounder and identifies β based on between-twin variation in test scores and birth weight. Logically, birth weight can vary due either to variation in gestation length, or to variation in fetal growth rates. By focusing on twins, we necessarily hold gestation length constant. Our estimates are identified, therefore, by variation in fetal growth rates. 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, this could present a source of bias. There are three different reasons why we might observe differential inclusion in the test score data related to poor neonatal health. First, parents may send one child to public school but the other to private school; if parents systematically send their heavier or lighter twin to different schooling settings, this could affect the observed relationship between birth weight and test scores, conditional on being in the public school setting. Second, since

[14] 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.

Florida exempts students from the FCAT in case of severe disability,[15] any relationship between birth weight and the likelihood of severe disability could affect our estimated relationship of interest. Third, if low birth weight children are more likely to miss the exam because of illness or absence, the effect on the estimates would be similar to the bias that results from differential disability.

That said, the evidence suggests that these potential internal validity concerns are not major issues. When we estimate twin fixed effect regression models in which the dependent variable is whether the twin ever attended public school (79.5 percent of the Florida twin birth population) the coefficient on log birth weight is -0.012 (with a standard error of 0.008). If the dependent variable is an indicator for beginning in public school by first grade (77.0 percent of twin births), the coefficient estimate is -0.007 (with a standard error of 0.008). If the dependent variable is whether we ever observe an FCAT score for the child (69.1 percent of twin births), the coefficient estimate is -0.003 (standard error of 0.009), and if the dependent variable is whether we observe an FCAT score in every possible year expected if the student did not leave Florida public schools (65.5 percent of twin births) the coefficient estimate is 0.006 (standard error of 0.004). In sum, it appears that relatively heavy and relatively light twins are remaining in public school and taking the FCAT at highly similar rates. These similarities diminish the potential internal validity concerns associated with differential test-taking rates.

IV. Preliminary results – heavier versus lighter twins

A. Test scores of heavier versus lighter twins

Before presenting the main regression results, we begin with simple comparisons of the test scores of heavier and lighter twins based on birth weight. These results, which aggregate twin pairs with small birth weight discordance with those with large birth weight discordance, are shown in figure 3. Nonetheless, they clearly demonstrate the first main result of the paper. Figure 3 shows the average within twin pair difference in test score between the higher birth weight twin and the lower birth weight twin, while figures A1 and A2 in the online appendix show the same patterns for mathematics and reading,

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

respectively. These figures show test score differences for the average of math and reading scores calculated separately at grades three through eight, along with the 95-percent confidence interval around those differences.[16][17] Note that these figures do not reflect panels of students, so there are different groups of children in each grade.

Within twin pairs, on average the heavier born scores about five percent of a standard deviation higher than the lighter born twin. This difference in test scores is statistically distinguishable from zero, and is stable from third through eighth grades, covering ages from approximately 9 to 14. This comparison holds constant any confounding factor that varies at the family, mother or birth level. The results imply that 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.

Figure 4 breaks down this mean difference by birth weight discordance[18]; the quartile with the most similar children (in terms of birth weight) averages just 2.5 percent discordance, while the quartile with the most different children averages 23.9 percent discordance. 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 shows that the upward-sloping relationship between birth weight discordance and test score differences is present and clear when we break down the twin pairs into fine discordance bins (one for each percentage point, and a final bin for twin pairs with greater than twenty percent discordance.) These findings foreshadow the main findings of this paper.

B. Testing for differential attrition

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

[17] 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.

[18] 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.

As described above, we observe test scores only for students who are in Florida public schools, and a small fraction of public school students miss the exam because of absence or because they have a profound disability.[19] Attrition from the testing data is only a concern for our estimates if students with missing test scores would have had particularly high or low scores relative to their twin, and since we are including twin-pair fixed effects in all of our models, attrition only causes bias if one twin leaves the testing data and the other remains. [20]

Such differential attrition or missing test score data is possible, so in addition to the twin fixed effects estimates described above, we present three tests of whether non-random attrition from the sample biases our estimates. Each of these tests indicates that the estimates are not meaningfully biased by non-random attrition. First, in appendix figure A4 we plot the difference in test scores between heavier and lighter birth weight twins, restricting the sample to those for whom we observe both twins for each of the six grades. The pattern is essentially unchanged from what we saw for the full sample in figure 3. The stability of the difference in test scores between heavier and lighter twins does not appear to be affected by changing selection out of the sample as twins age.

Second, we measure directly the amount of differential attrition from the sample between third and eighth grade. Starting with the sample of twins we first observe in third grade, appendix figure A5 shows the difference in the fraction of heavier and lighter birth weight twins tested in each subsequent grade. The figure shows that lighter birth weight twins are slightly more likely to have missing tests in the sixth through eighth grades, possibly because they are pulled from public schools and possibly because they are still enrolled in public school but missed the exam.[21] However, the magnitude of the

[19] While over 30 percent of Florida twins receive some special education services, compared to 12 percent of singletons, a large majority of students with disabilities in Florida take the FCAT. Only students with disabilities such as severe mental retardation or severe autism are exempt from the FCAT.

[20] There has been a secular trend toward more very low birth weight infants surviving and entering the educational system impaired. See, e.g., Zwicker and Harris (2008) and Aarboudse-Moens et al (2009).

[21] We start off with a sample of twin pairs with twins old enough for 3rd grade. We have 2.9, 2.5, 2.3, 2.5, 1.9 and 1.8 percent of pairs in grades 3, 4, 5, 6, 7 and 8 respectively where twins are old enough to be in the grade but neither of them is tested. We have 2.4, 2.2, 2.3, 2.5, 2.7 and 2.6 percent of pairs in grades 3, 4, 5, 6, 7 and 8 respectively where twins are old enough for the grade but we observe test scores for only one twin. We have already demonstrated above that in cases in which only one twin is missing a test score, it is not systematically the case that the twin with the missing score is the lighter twin.

difference is not large enough to significantly affect the relative magnitude of the within twin pair test score differences, a conjecture also we test directly below.

Third, we can put bounds on the magnitude of any bias resulting from differential attrition of high and low birth weight twins between third and eighth grade. Figure A6 shows three sets of estimated differences by grade -- one where we replaced missing test scores with the 5th percentile of test scores in that grade, another where we replaced missing test scores with the 95th percentile, and a third where we replaced missing test scores with the imputed conditional mean. As the figure shows, even assuming that students who no longer had an FCAT score after third grade had extremely low or extremely high test scores does not change the conclusion that the within twin pair difference in test scores is remarkably stable from third through eighth grade. Taken together, the results show that attrition patterns do not significantly affect the patterns of results between third and eighth grade.

V. Main results

A. Pooled results for full sample

We now turn to our main regression results. As described above, the basic regression model is an ordinary least squares estimate that includes twin-pair fixed effects, a gender dummy, and a dummy for within-twin-pair birth order. The dependent variable is the standardized FCAT score, and the regressor of interest is the natural logarithm of birth weight in grams. We report some results based on separate regressions for each grade from three through eight, and other results that pool test scores across all six grades. In the pooled regressions, standard errors are clustered at the individual level (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.[22]

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

[22] An earlier version of this paper (Figlio, Guryan, Karbownik and Roth, 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.

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 figure also shows clearly that the shape of the relationship between test scores and birth weight is very similar whether or not we condition on twin-pair fixed effects.

We present our main result in column 2 of table 2.[23] The results show that within twin-pairs higher birth weight is indeed associated with higher test scores in grades three through eight. The estimated coefficient of 0.441[24] (for combined reading and mathematics) implies that a ten percent increase in birth weight is associated with just under one-twentieth of one standard deviation increase in test scores.[25] The coefficient is precisely estimated, with a *t*-statistic of 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. The results are somewhat larger for mathematics than for reading, but the patterns are the same for both subjects; therefore, for ease of presentation, from here on we concentrate on the combined math and reading results.[26]

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

[23] In this paper we treat each child-test combination as a separate observation, and cluster the standard errors. Our results are fundamentally unchanged if we instead collapse all observations on a child to a single observation and weight each child equally, regardless of the number of years of data observed.

[24] If we weight observations by one over the number of observations seen for an individual, all results are nearly identical. For instance, the coefficient on log birth weight in the pooled combined math and reading twin fixed effects model is 0.447, rather than the unweighted coefficient of 0.441.

[25] 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 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.

[26] 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 1 percent of a standard deviation higher average reading and math scores.

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.[27] Combining this with our finding that a 1000 gram increase in birth weight is associated with a 0.186 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.

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 the test scores of heavier and lighter born twins stratified by mother's education. The figure clearly shows that the difference in test scores resulting from differences in birth weight is small compared with differences in achievement associated with mother's education. Each of the differences between heavier and lighter born twins shown in the figure is statistically significant. However, it is clear that in terms of math and reading achievement, it is better to be the lighter born twin of a college educated mother than the heavier born 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.[28]

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

[27] 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.

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

grow larger as children age, or is it present by age 9 and does it remain constant through the upper elementary and middle grades? The structure of the data allows us to estimate the effect of birth weight on test scores separately by grade to address these questions.

The results are presented in columns 3-8 in table 2. The table shows the estimated effect of log birth weight from twin fixed effects models that are estimated separately for test scores from each grade, three through eight. As is the case throughout the paper, grade refers to the "imputed" grade twin pairs would have attended if they had progressed on a normal schedule after we first observe their third grade exam score, and as mentioned above, our results are not sensitive to how we define grade.

The table shows that the effect of birth weight on cognitive achievement is in fact present by the third grade. The twin fixed effect estimate of the effect of log birth weight on test scores in third grade is 0.442. The grade-specific estimated effect remains fairly stable from third through eighth grade, ranging from 0.372 to 0.524. Note that while the *F*-test that the grade-level estimated effects are identical is rejected at a moderate level of statistical significance (p=0.067), there is no evidence that this effect follows a substantial systematic pattern as children age. In a regression model in which we interact the log of birth weight linearly with grade in school, the coefficient estimate on the interaction term is one-two thousandth the magnitude of the coefficient on the log of birth weight. These results suggest that the effects of neonatal health do not substantially change, between ages 9 and 14. Rather, 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, Guryan, Karbownik and Roth, 2013), we look even further back, to the beginning of formal schooling.[29] In various years between 1998 and 2008, Florida performed universal kindergarten readiness screening and recorded this screening in its Education Data Warehouse. From 1998 through 2001 all kindergarten entrants were screened with the School Readiness Checklist (SRC), a list

[29] 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.

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.[30] 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 84.0 percent of children were deemed either above average or low risk according to the DIBELS. Making use of twin comparisons in a linear probability model[31], 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. [32] All estimates are statistically distinct from zero at conventional levels. These results make it clear that the effect of neonatal health on cognitive development that we document for ages 9-14 is present by age 5.

C. Role of genetic differences between twins

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

[30] For more details about the structure and interpretation of DIBELS, see, e.g., Hoffman et al. (2009).

[31] We present marginal effects from a linear probability model because they are the most transparent to interpret. The pattern of results and statistical significance is extremely similar when we instead estimate conditional logit models.

[32] In Figlio, Guryan, Karbownik and Roth (2013) we go into detail about the metrics one can employ to directly compare the dichotomous kindergarten readiness assessments to later continuous test scores.

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. The fourth and fifth rows of table 2 show estimates separately for same-sex and different-sex twins.

As can be seen in the table, the estimated effect of birth weight is extremely similar for same-sex twins (0.448) and different-sex twins (0.421), suggesting that the estimated relationship is within the same general range regardless of zygosity. This result is consistent with results reported in BDS, who find no significant differences in the effect of birth weight on adult earnings between same-sex and opposite-sex twins, nor do they find significant difference in estimated effect of birth weight on earnings for monozygotic twins and dizygotic same-sex twins in their sample with available zygosity information. Taken together, the results suggest that genetic differences between twins are unlikely to be driving a large portion of the relationship between birth weight and later outcomes.

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 considerable cost in terms of external validity. Twins have older and more educated mothers, for instance, and they weigh considerably less on average 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 per se 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, age and number of prior births.

The sixth 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 age, 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 twins 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 seventh 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 realized birth weight.

Singletons include some babies whose birth weight is high enough that it likely indicates an underlying poor 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 0.441. 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 final 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.

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 gestational 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. The left panel of the figure pertains to twins, while the right panel of the figure pertains to singletons. This figure makes it apparent 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

Two special features of the Florida context allow us to investigate heterogeneity in the effects of birth weight in ways that have not been possible in other related work to this point. Florida has a remarkably heterogeneous population. 22.6 percent of all births in Florida are to black mothers, 23.0 percent of births are to Hispanic mothers, and 23.5 percent to foreign-born mothers. Florida has a wide income distribution and a substantial distribution in terms of maternal education. The diversity of demographics in the state, combined with the size of the dataset make comparisons of birth weight effects across racial, ethnic and socio-economic groups possible.

It is inherently interesting to learn about whether the long-term effects of *in utero* conditions on cognitive development vary across demographic and socio-economic groups. Moreover, examining heterogeneity in the effects may shed light on the mechanisms by which neonatal health affects cognitive skills. There are significant differences in household income, wealth and education by race, ethnicity and immigrant status. If these factors, each of which is strongly correlated with student achievement at the population level, are substitutes with neonatal health in the production of cognitive skills then we should expect to see larger effects of birth weight on test scores for more disadvantaged groups. If income, wealth and parental education are complements with neonatal health, one would expect to see larger effects for more advantaged groups.

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 appendix table A1 we report grade-by-grade results for all subgroups of the twins analysis. The first and second columns of table 3 report the group mean test score for twins and singletons, respectively, in each subgroup, while the third and fourth columns report the mean and standard deviation of each group's birth weight for twins and singletons. The fifth column reports the mean and standard error of the estimated effect of birth weight on pooled third through eighth grade test scores in a twin fixed effects model. The sixth through eighth columns report the parallel findings for singletons: The estimated coefficient on log birth weight (column 6), log birth weight conditional on gestation length (column 7)[33], and gestation length (column 8).

Before turning to the heterogeneity analyses, we first consider whether the estimated effect of log birth weight varies by gender. As can be seen in the first panel of table 3, the results are very similar for boys and girls.[34] 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 virtually identical (0.452 and 0.445, respectively). The same is true when we compare OLS estimates for the singleton population (0.296 versus 0.276), or the OLS estimates conditional on gestation (0.440 versus 0.407). Likewise, the estimated OLS relationship between weeks of gestation and scores is 0.013 for both boys and girls.

The second panel of table 3 stratifies births based on whether the mother has a medical history that could pose a problem for the pregnancy or delivery.[35] Around one-quarter of mothers have one of these risk factors. We observe that the pooled fixed effects estimates are very similar (0.415 for mothers with medical history, and 0.448 for mothers without medical history), as are the log birth weight coefficients for singletons (for instance, 0.373 for mothers with medical history, and 0.431 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 shows 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). This represents a massive range of student advantage, with average group test scores as low as -0.477 and as high as 0.663, reflecting gaps that are

[33] In the singleton 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.

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

[35] 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.

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 0.359 and 0.527. The OLS coefficient estimates in the singleton population range from 0.251 to 0.326, and the OLS coefficient estimates on birth weight conditional on gestation range between 0.344 and 0.490. 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.[36] 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.

If we find that children from higher socio-economic status families tend to have a smaller relationship between log birth weight and student outcomes, this finding could indicate that parental inputs are substitutes for poor neonatal health. If instead we observe that high socio-economic status families have larger relationships between log birth weight and student outcomes, this observation would suggest that neonatal health and family inputs may be complements. Indeed, there is some reason to believe that this latter possibility is the case: In the twin fixed effects specifications, the estimated effect of log birth weight is increasing in maternal education, is higher for families living in medium and high-income neighborhoods than for families living in low-income neighborhoods, and is higher for children with married parents than for those with unmarried parents.

[36] 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.

However, the relative patterns are not the same for singletons, and the subgroups (e.g., those split by maternal education and neighborhood income) are not mutually exclusive, suggesting that these relative rankings of coefficient estimates may just be chance occurrences.

To explore this question more fully, we pursue an approach similar to that employed by Hoynes, Miller and Simon (2012) 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 student gender, maternal race, maternal ethnicity, maternal immigrant status, maternal marital status, maternal age, maternal education, and neighborhood income to predict student test scores, and then divide the students into ten mutually-exclusive groups; these groups range in mean predicted test scores from -0.680 to 0.790 in the twins population – a range greater than a full individual-level standard deviation of the test score distribution.[37] Figure 8 plots each group's estimated coefficient on log birth weight against the group's mean score. We do this for two different models – the twin fixed effects model and the comparable OLS model for singletons: conditional on gestation and restricted to the population of singletons whose birth weights fall within the observed range of twin birth weights.

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 0.29 and 0.67 in the twin fixed effects model, and between 0.29 and 0.48 in the singletons OLS model, and the estimated effects are both statistically and economically significant for every demographic and socio-economic group analyzed.[38] 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 test scores estimated by BDS.

The second pattern the figure demonstrates is that there does appear to be a modest upward-sloping relationship between estimated treatment effects and the

[37] The groups range in mean test scores from -0.603 to 0.737 in the case of singletons.

[38] We have also estimated specifications in which we interact log birth weight separately with all of the 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 0.19 to 0.63. Appendix figure A7 plots the estimated marginal effects of log birth weight for the full distribution of possibilities in this specification.

subgroup's mean test score. This positive relationship indicates that the effects of birth weight are somewhat larger for relatively advantaged groups of children than they are for relatively disadvantaged groups of children. The slopes of the lines plotted in figure 8 are 0.019, with a standard error of 0.014, in the case of the twin fixed effects model, and 0.012, with a standard error of 0.003, in the case of the singletons OLS model.[39] The two lines are very similar in terms of both slope and intercept. Therefore, while by no means definitive, these patterns indicate that poor neonatal health may modestly disproportionately affect children growing up in high socio-economic status families, and are suggestive that neonatal health and parental resources are complementary.[40]

VI. 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. There is great attention paid by public health officials and medical practitioners on the thresholds of 1500g and 2500g, the conventional delimiters of very low birth weight and low birth weight, respectively.

[39] We estimate the standard errors of the slopes of these lines by bootstrapping. We randomly drew twin 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 (singleton) models for each of the ten deciles. For both twins and singletons, we ran 1000 replications of these 10-observation regressions and calculated the standard deviation from these slopes for our bootstrapped standard errors.

[40] 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 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; this is the age breakdown that Bitler uses to proxy for IVF/ART likelihood. Next, we conduct the identical analysis for twins who were the first children born to the mother to those who were not the first children born to the mother, given that IVF/ART is more likely amongst families with previous fertility challenges. 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.

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 babies.

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 lighter-born twin's birth weight,[41] 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.[42]

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).[43] 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 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 table A2 break down our estimates into different groups based on the birth weights of the twins. As can be seen, the estimated effect of a marginal increase in birth weight is quite similar for low birth weight (<2500g) and normal birth weight (≥2500g) children;

[41] 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.

[42] 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: 0.865).

[43] At very small discordances of less than three or four percent, the estimates are far 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.

the estimate for low birth weight twin pairs is 0.473, and for normal birth weight twin pairs it is 0.529, and the two pooled coefficients are not statistically distinguishable (p-value: 0.632). Likewise, the estimated effects reported in rows 4 and 5 of the table for very low birth weight (<1500g) and 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, 0.589, 0.518 and 0.529. An *F*-test fails to reject that these three estimates are the same (p-value: 0.918).[44]

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

The results presented thus far have demonstrated that there is a robust relationship between birth weight and third through eighth grade test scores, and that this relationship is remarkably stable as children age through preadolescence, across different demographic groups, and across different socio-economic groups. The stability of this relationship is all the more notable because the marginal effect of birth weight does not vary 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

[44] We test more formally the assumption suggested by our non-parametric estimates that the linear in log birth weight specification is reasonable. The sixth row reports the result from a regression that replaces the log of birth weight with birth weight in thousands of grams, but which is otherwise equivalent to the baseline specification. We estimate that a marginal increase of 100g of birth weight is associated with about 0.02 standard deviations higher pooled test scores. In another specification check (row 7), we interact birth weight in grams with the average of the twin pair's birth weight. (There are two coefficients reported in this row, the coefficient on birth weight and the coefficient on its interaction with the deviation from mean twin pair birth weight.) Here, the twin pair average birth weight is demeaned from the sample average so that the birth weight coefficient represents the marginal effect of birth weight in a twin pair of average birth weight. The results show that the marginal effect of a gram of birth weight is smaller in heavier twin pairs, as indicated by the negative and significant coefficient on the interaction term. This result is consistent with the linear-in-logs model, in which test scores are proportionally related to birth weight. Based on these further results, we conclude that the linear-in-logs specification is a good approximation of the relationship between birth weight and cognitive skills.

to his counterpart if the twin pair attends a high quality 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., 2011a, Chetty et al., 2011b) 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. In addition, we stratify schools based on average proficiency levels and average student gains from year to year. While these are only three of a wide range of ways in which one could evaluate school quality, they are sufficiently different[45] that similar findings across the three measures would provide strong evidence of the potential effects of school quality in ameliorating or exacerbating the relationships between birth weight and student cognitive development. In our analysis, therefore, we measure school quality using (1) the state awarded letter grade, (2) the school's average FCAT proficiency level during our sample period, and (3) the school's average year-to-year student FCAT gain score over our sample period. In addition, we have coded, to the closest degree possible in our data, three other state/city school grading systems that have received considerable media attention – the systems in Indiana, Louisiana, and New York City.[46] While these rating systems are all still based largely or exclusively on standardized test scores, they rank order schools considerably differently – New York City's ranking of Florida schools and Louisiana's ranking of Florida schools are

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

[46] Indiana grades schools based on a combination of average test scores, percentage of low-performing (relative to the school) students making high growth, percentage of higher-performing students making high growth, and percentage of students making low growth. Louisiana grades schools based on a nonlinear transformation of average performance levels of students. New York City grades schools on a combination of average performance levels, percentage of students making at least a year's worth of progress, percentage of low-performing students (relative to a school) making at least a year's worth of progress, average change in proficiency for low-scoring (relative to the state) students, average change in proficiency for higher-scoring students, and measures of school environmental factors.

correlated at 0.40, and New York City's ranking and Indiana's ranking correlate at 0.60. Therefore, while we are constrained to measure school quality mainly using student test scores, we are considering a wide range of school quality measures, including those that closely resemble the full set of metrics currently being used by every U.S. state, with a combined population of 88 million residents, that assigns letter grades to schools.[47]

The results of the school quality analyses are presented in tables 4 and 5. The first panel of table 4 shows estimates separately for twins who attended schools that received an A, a B, and a C or below.[48] 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 (0.500). The smallest estimated effect is for twins attending A schools (0.405), 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 0.274 and 0.284 and the estimated coefficient on birth weight conditional on gestational length ranges from 0.378 to 0.413.

The second panel in table 4 presents results where school quality is measured based on the school's average FCAT scores. About 65 percent of twins attended schools with scores that are above the state median average score – unsurprising given that families of twins are disproportionately older, more educated, and live in neighborhoods with higher median income. Though average test scores for both twins and singletons are certainly different in high- and low-average-test-score schools, the estimated effect of

[47] In addition to Florida, Indiana, Louisiana, and New York City, the other states assigning letter grades to schools at the time of writing are Alabama, Arizona, Mississippi, New Mexico, North Carolina, Ohio, Oklahoma, South Carolina, and Utah.

[48] We combine C, D, and F-graded schools in this analysis because highly educated and older families, who are more likely to have twins, are more likely to live in "better" [i.e., higher graded] school zones than the general population, and because the state of Florida has awarded relatively few grades of D and F. In the overall population, 5.0% and 0.8% students attend D and F schools respectively, while among twins these rates are 3.4% and 0.6% respectively.

birth weight does not vary meaningfully for either group. We estimate that the marginal effect of log birth weight for twins attending schools with above-median FCAT scores is 0.423. For twins attending schools with below-median FCAT scores, we estimate the effect to be 0.438. The corresponding figures for singletons are similar: 0.404 and 0.395 when conditioning on gestational length and 0.267 and 0.282 when not.

Our estimates of the effect of log birth weight on test scores also does not vary between schools with above and below average FCAT gains. These estimates are shown in the third panel of table 4. We estimate that the marginal effect of log birth weight on test scores for twins attending a school that had below-median year-to-year gains in FCAT scores is 0.448. For twins attending a school that had above-median FCAT gains, we estimate the marginal effect of log birth weight to be 0.433. Again, the corresponding figures for singletons are extremely close: 0.429 and 0.413 when conditioning on gestational length and 0.286 and 0.285 when not.

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 New York City's, Louisiana's, and Indiana's grading systems into thirds[49] 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 almost never 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 or from a singletons model controlling for gestational length or from a singletons model without controlling for gestation. Third, the only cases in which there is a monotonic relationship between the point estimates on birth weight and measures of school quality are for singletons in the New York City grading metric specifications, and there the coefficient estimates are all extremely similar – the most similar of all the comparisons.[50]

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

[49] We do not calculate the actual school letter grades that would be inferred from the different states' rankings because Florida's test scores and the other states' test scores are not directly comparable.

[50] 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.

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 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 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.[51]

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 1992, 1996, Krueger & Whitmore, 2001, Chetty et al., 2011a, Chetty et al., 2011b) of schooling have been shown to have significant positive impacts on earnings and other outcomes. If attending a better school does not completely remediate the effects of early health deficits on cognitive development, maybe schools currently lack the resources 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

[51] 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.

view is supported by a few observations. The difference in birth weights (or cognitive capacities) between twins is probably far more noticeable to parents than to classroom teachers. To parents a 15 percent difference in twins' birth weight would be noticeable, but to a teacher nine to fourteen years later twins' initial birth weight would be insignificant compared to the variation she observes in the classroom. Even twins with large discordance in birth weight and with the resulting differences in cognitive achievement probably appear to the teacher to be 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, for that matter, similar children under a different dimension – very 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, at least not systematically.

VIII. 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: Twins with higher birth weight enter school with a cognitive advantage that appears to remain stable through the elementary and middle school years. The patterns 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, Guryan, Karbownik and Roth, 2013), and even as they are in very early childhood (Hart, 2008).[52] 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 et al. (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 discordance greater than 20 percent) parents "redshirt" one twin but not the other – starting twins in school at different ages.[53]

It is the case that 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 of families – and,

[52] 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.

[53] 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).

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. The rate of induced births has increased considerably in recent decades in the United States; between 1990 and 2008, the induction rate more than doubled from 9.6 percent to 23.1 percent, and has been especially pronounced for infants of at least 37 weeks of gestation. Our findings of a positive relationship between birth weight and test scores throughout the birth weight distribution – as well as a positive reduced-form relationship between gestational length and test scores – suggest that the costs of early induction may be higher than some parents-to-be and physicians think. On the education side, our findings suggest that there may be educational benefits to greater communication between schools and health care providers, as doing so may help schools to target their resources more effectively. These last points are speculative, and are the topics of our current research agenda.

Bibliograpy

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

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

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

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

Bharadwaj, Prashant, Juan Eberhard, and Christopher Neilson. 2013. "Health at Birth, Parental Investments and Academic Outcomes", working paper, University of California-San Diego

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

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

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

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. 2011a. "How Does Your Kindergarten Classroom

Affect Your Earnings? Evidence from Project Star", *Quarterly Journal of Economics* 126(4): 1593-1660

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

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", NBER 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", NBER 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", working paper, Northwestern University

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

Hoynes, Hilary W., Douglas L. Miller and David Simon. 2012. "Income, the Earned Income Tax Credit, and Infant Health", NBER working paper # 18206

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", working paper, Duke University

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, Waldemar A. Carlo. 2013. "Extremely Low Birth Weight and Infant Mortality Rates in the United States", *Pediatrics* 131: 855-60

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

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

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

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

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

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

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

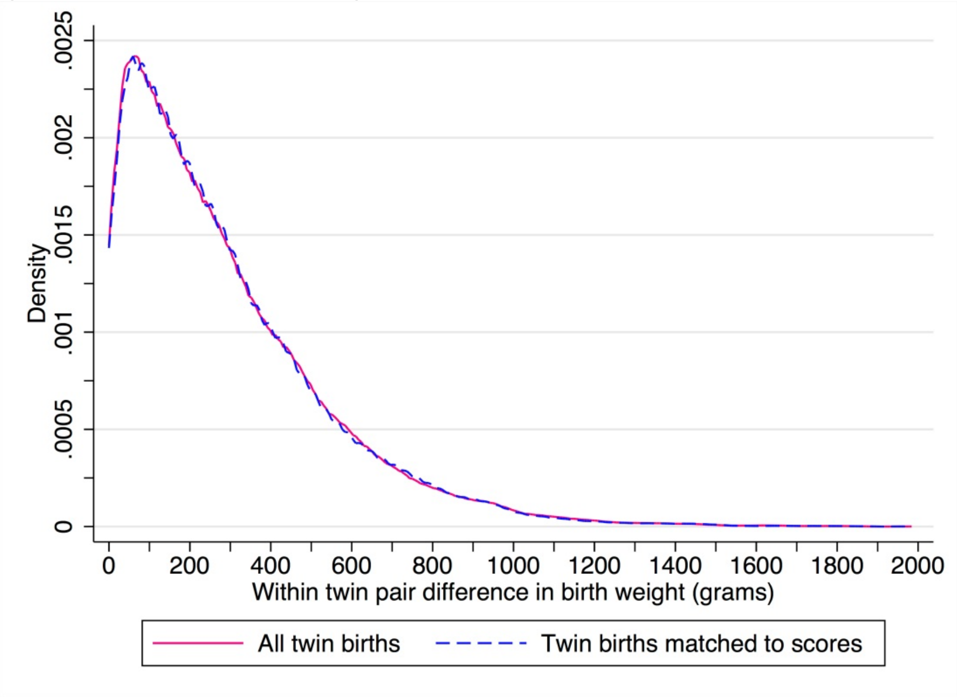
Sacerdote, Bruce. 2001. “Peer Effects with Random Assignment: Results for Dartmouth Roommates”, *Quarterly Journal of Economics* 116(2): 681-704

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

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

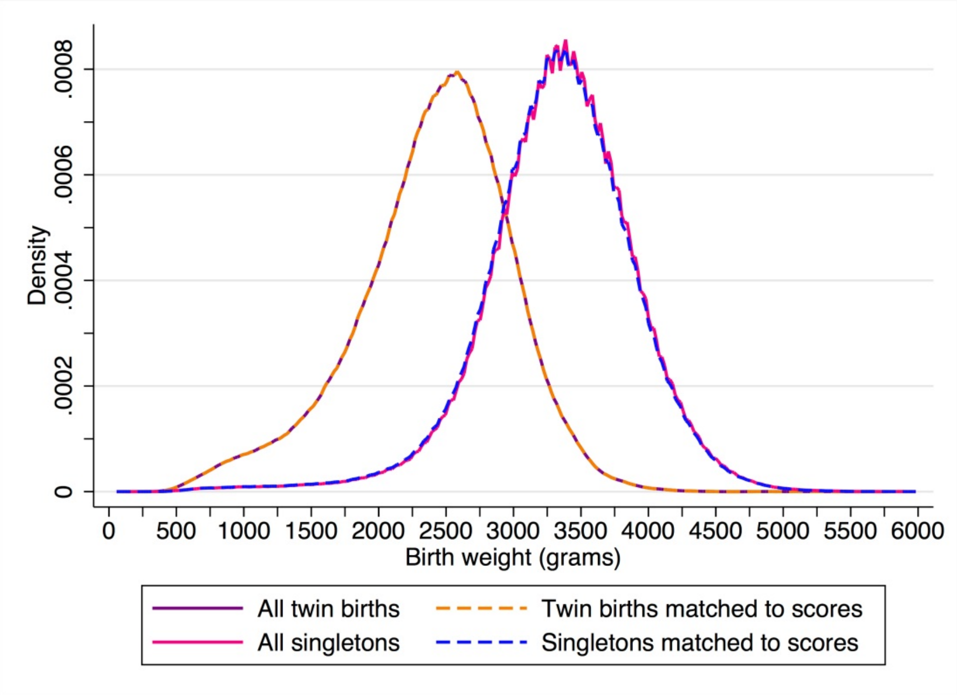
FIGURES

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



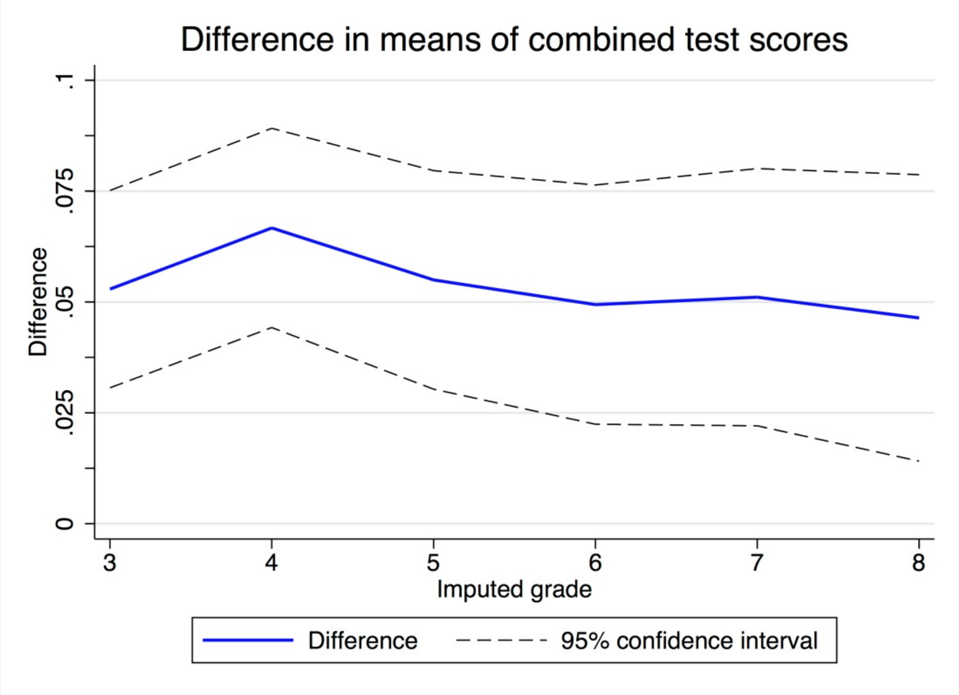
Note: Figure 1 plots kernel density distributions of within-twin-pair difference in birth weight for all twin births in Florida (solid pink line) between 1992 and 2002 and twin births who were born in Florida and were successfully matched to Florida public school records (dashed blue line). Distributions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

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



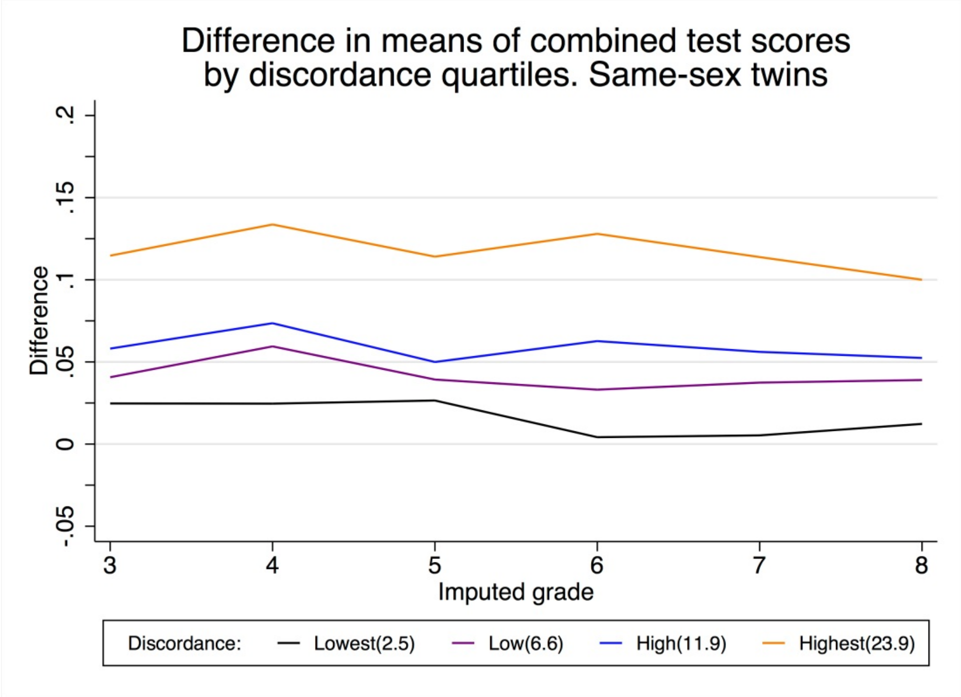
Note: Figure 2 plots kernel density distributions of infant birth weight for all singletons (solid pink line) and twins (solid purple line) born in Florida between 1992 and 2002 as well as infant birth weight distribution of singletons (dashed blue line) and twins (dashed orange line) that were successfully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

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



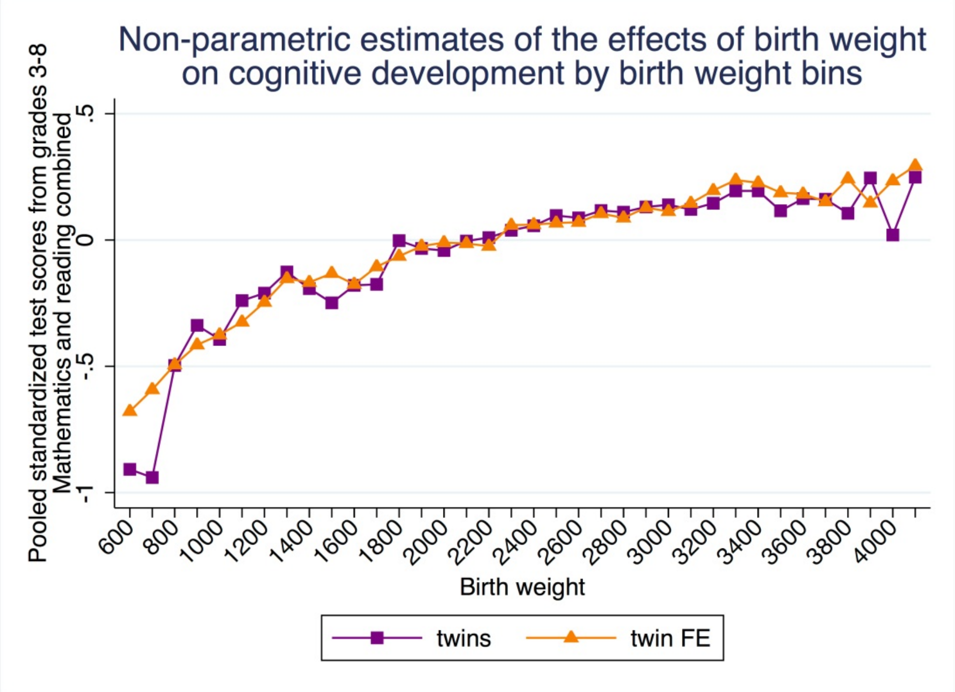
Note: Figure 3 plots difference between the mean test score of heavier and lighter twin from each pair in each grade and the respective 95% confidence interval of this difference. Mean test score is constructed as an average of scores in mathematics and reading for each individual in each grade where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

Figure 4: Means of scores by discordance quartiles



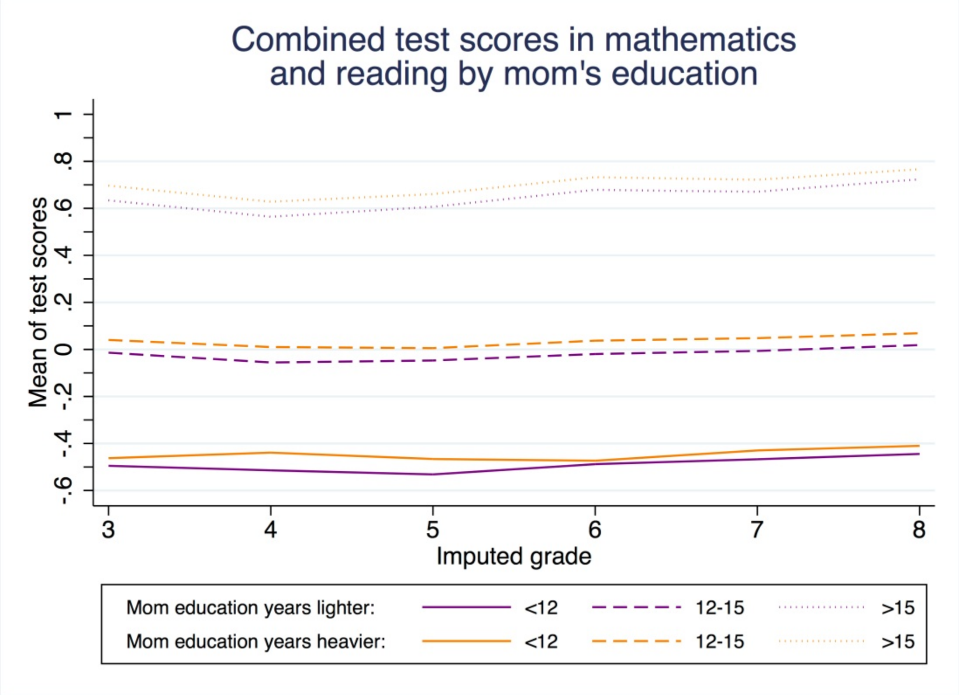
Note: Figure 4 plots difference between the mean test score of heavier and lighter twin from each pair in each grade for four quartiles of discordance in birth weight. Mean test score is constructed as an average of scores in mathematics and reading for each individual in each grade where we observe 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



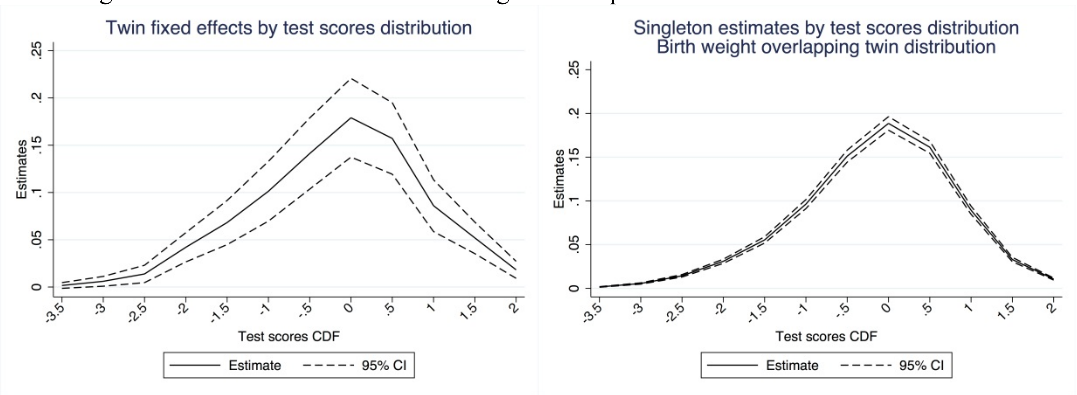
Note: Figure 5 plots coefficients from OLS (purple solid line) and twin-FE (orange solid line) models where the dependent variable is the mean of pooled grades three through eight of combined mathematics and reading test scores for each individual and the independent variables are indicators for 37 weight 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



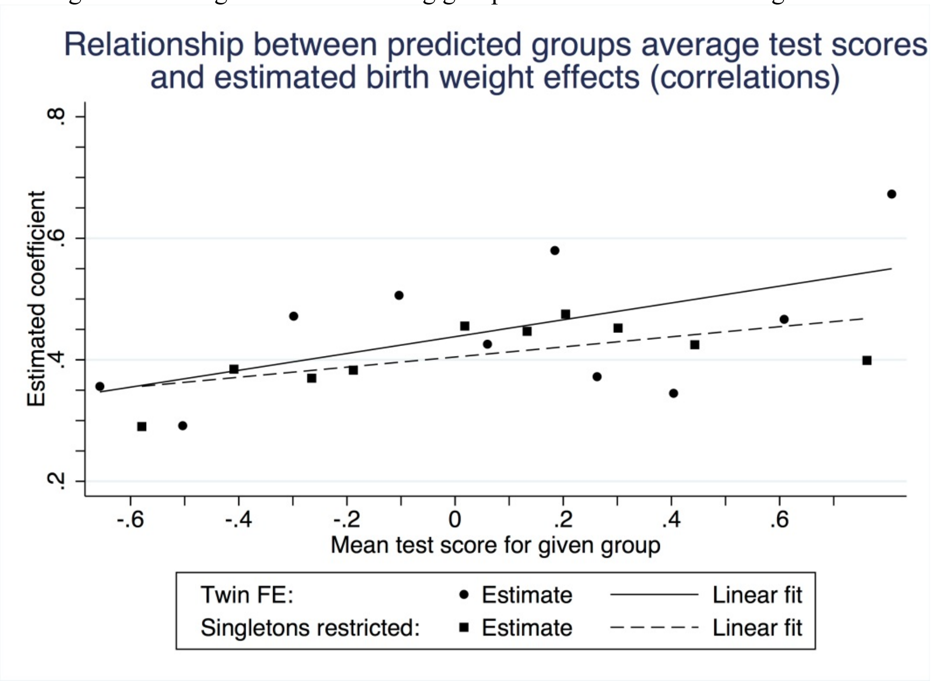
Note: Figure 6 plots means of combined mathematics and reading test scores for lighter and heavier twins from each pair stratified by maternal education. Purple lines correspond to averages for lighter while orange lines correspond to heavier twins. Solid lines present means for high school drop-out 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



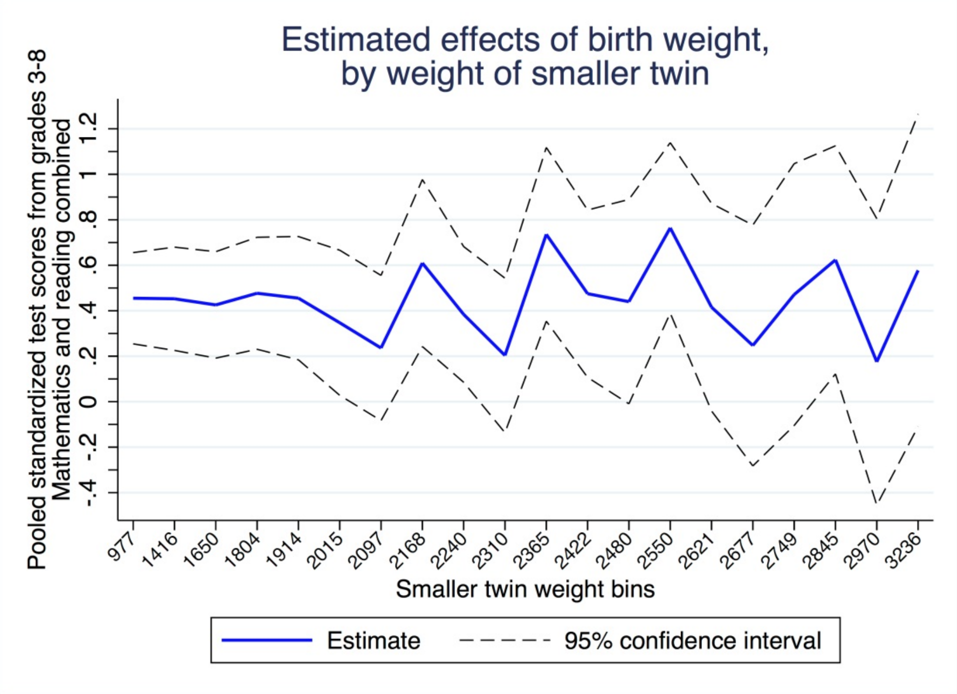
Note: The left panel of figure 7 plots coefficients on log birth weight from a standard twin FE regressions estimated separately for CDF (greater than -3.5, greater than -3, etc.) bins of test scores distribution. The right panel of figure 7 plots coefficients on log birth weight from a standard singletons regressions conditional on gestation and using birth weight overlapping twin birth weight distribution estimated separately for CDF bins of test scores distribution.

Figure 8. Average test scores among groups and estimated birth weight effects



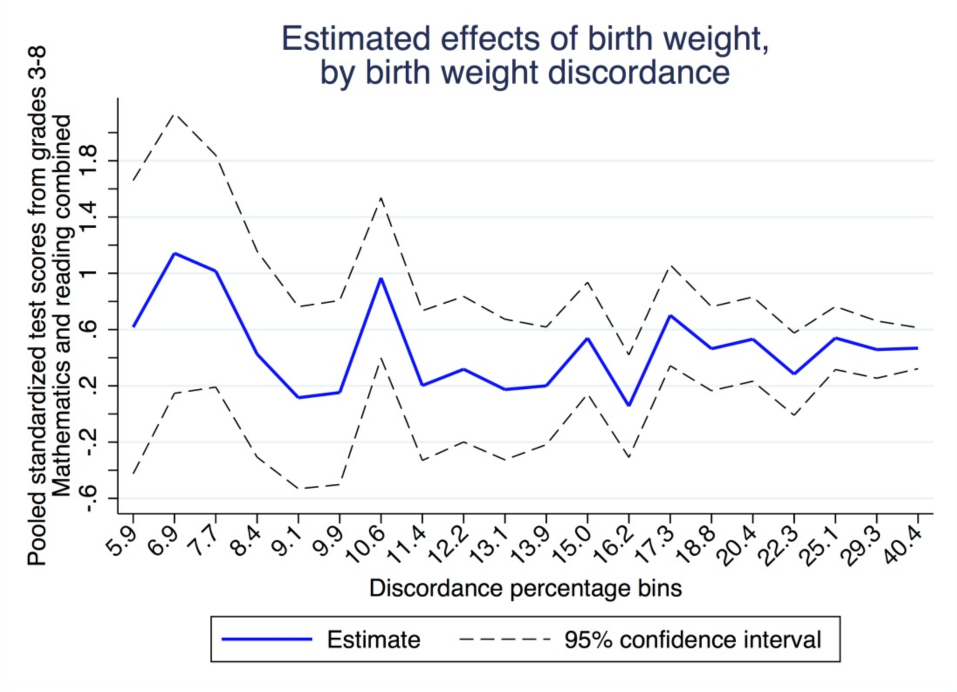
Note: Figure 8 plots the estimates for the 10 predicted groups based on the regression of test scores on gender, race, ethnicity, immigrant origin, marital status, maternal education, maternal age categories and income indicators corresponding to those from table 3. These groups are not overlapping. In this graph income from 1992 and 1993 is imputed based on observables. Groups are calculated only for twin pairs with all information available and for all singletons with birth weight in a range of 847 to 3600 grams.

Figure 9: Estimated effects of birth weight, by weight of smaller twin



Note: 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 birth weight discordance in each of the 20 bins.

Figure 10: Estimated effects of birth weight, by birth weight discordance



Note: 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 twin population

	(1)	(2)	(3)	(4)
Maternal attribute	Full population of births	Population of kids matched to Florida school records	Population of kids with a third-grade test score	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	20.5	17.5	16.2	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	23.5	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	1,334,008	28,466

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

Table 2: Estimated effects of birth weight on cognitive development

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Pooled		Imputed grade: Twin FE models					
	OLS	Twin FE	3	4	5	6	7	8
				Twins				
	Mathematics:							
Ln(birth weight)	0.353*** (0.024) [126,212]	0.497*** (0.042)	0.475*** (0.051) [28,398]	0.579*** (0.051) [26,466]	0.535*** (0.052) [22,936]	0.490*** (0.061) [19,294]	0.411*** (0.067) [16,106]	0.427*** (0.070) [13,012]
	Reading:							
Ln(birth weight)	0.216*** (0.023) [126,384]	0.391*** (0.041)	0.415*** (0.048) [28,374]	0.464*** (0.051) [26,436]	0.326*** (0.055) [22,928]	0.373*** (0.059) [19,312]	0.370*** (0.062) [16,180]	0.348*** (0.069) [13,154]
	Average of mathematics and reading:							
Ln(birth weight)	0.285*** (0.022) [126,826]	0.441*** (0.039)	0.442*** (0.043) [28,466]	0.524*** (0.045) [26,542]	0.430*** (0.047) [23,006]	0.426*** (0.053) [19,370]	0.387*** (0.056) [16,216]	0.372*** (0.061) [13,226]
	Average of mathematics and reading – same-sex twins:							
Ln(birth weight)	0.300*** (0.027) [87,378]	0.448*** (0.043)	0.461*** (0.050) [19,410]	0.527*** (0.053) [18,166]	0.409*** (0.053) [15,852]	0.466*** (0.059) [13,392]	0.395*** (0.062) [11,278]	0.362*** (0.067) [9,280]
	Average of mathematics and reading – opposite-sex twins:							
Ln(birth weight)	0.260*** (0.038) [39,448]	0.421*** (0.082)	0.393*** (0.086) [9,056]	0.511*** (0.088) [8,376]	0.478*** (0.097) [7,154]	0.329*** (0.112) [5,978]	0.365*** (0.122) [4,938]	0.394*** (0.135) [3,946]
				Singletons (*Average of mathematics and reading)*				
Ln(birth weight)	0.285*** (0.004) [5,757,145]	-	0.305*** (0.004) [1,255,620]	0.289*** (0.004) [1,182,390]	0.292*** (0.004) [1,041,614]	0.281*** (0.005) [889,644]	0.262*** (0.005) [757,170]	0.261*** (0.005) [630,707]
Ln(birth weight) | gestation weeks	0.332*** (0.005) [5,757,145]	-	0.346*** (0.005) [1,255,620]	0.336*** (0.005) [1,182,390]	0.337*** (0.006) [1,041,614]	0.329*** (0.006) [889,644]	0.313*** (0.007) [757,170]	0.316*** (0.007) [630,707]
Ln(birth weight) | gestation weeks [overlapping]	0.421*** (0.007) [4,029,066]	-	0.430*** (0.008) [883,301]	0.424*** (0.008) [829,941]	0.428*** (0.009) [728,842]	0.421*** (0.009) [621,098]	0.400*** (0.010) [527,542]	0.407*** (0.011) [438,342]
Gestation weeks	0.013*** (0.000) [5,757,145]	-	0.015*** (0.000) [1,255,620]	0.013*** (0.000) [1,182,390]	0.013*** (0.000) [1,041,614]	0.012*** (0.000) [889,644]	0.011*** (0.000) [757,170]	0.010*** (0.001) [630,707]

Note: Columns (1) and (2) present pooled grade three through eight results for OLS and twin-FE models. Columns (3) to (8) present OLS and twin-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 and only twin pairs where both twins 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 rage 847 to 3600 grams. This restriction provides overlapping distribution of birth weight among twins and singletons. The dependent variables are score in mathematics, score in reading and an average test scores in mathematics and reading. If the test score in mathematics is not available then reading is included and vice versa. The main variable of interest is natural logarithm of birth weight. The remaining independent variables in twin-FE models include infant gender 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). 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.

Table 3: Estimated effects of birth weight on cognitive development by child and mother characteristics

Characteristic	Sample	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Mean test score		Mean (SD) birth weight		Pooled twin FE estimate	Singletons		
		Twins	Singletons	Twins	Singletons		Birth weight	Birth weight | gestation	Gestation
(1) Gender	Boys	0.048	0.050	2473	3397	0.452***	0.296***	0.440***	0.013***
				(571)	(562)	(0.068)	(0.005)	(0.011)	(0.001)
	Girls	0.100	0.119	2369	3274	0.445***	0.276***	0.407***	0.013***
				(555)	(541)	(0.053)	(0.005)	(0.010)	(0.000)
(2) Maternal medical history	No medical problems	0.075	0.098	2457	3359	0.448***	0.300***	0.431***	0.011***
				(553)	(526)	(0.048)	(0.005)	(0.009)	(0.000)
	Medical problems	0.073	0.041	2356	3259	0.415***	0.251***	0.373***	0.015***
				(580)	(638)	(0.066)	(0.006)	(0.013)	(0.001)
(3) Maternal race	White	0.256	0.228	2457	3393	0.464***	0.293***	0.457***	0.011***
				(554)	(538)	(0.045)	(0.005)	(0.009)	(0.000)
	Black	-0.466	-0.363	2319	3180	0.387***	0.262***	0.344***	0.015***
				(585)	(576)	(0.082)	(0.006)	(0.013)	(0.001)
(4) Maternal ethnicity	Non-Hispanic	0.098	0.110	2413	3333	0.433***	0.283***	0.426***	0.012***
				(565)	(560)	(0.044)	(0.004)	(0.008)	(0.000)
	Hispanic	-0.035	0.001	2455	3346	0.481***	0.271***	0.385***	0.012***
				(564)	(539)	(0.079)	(0.008)	(0.015)	(0.001)
(5) Maternal immigration history	Non-immigrant	0.073	0.078	2413	3334	0.439***	0.284***	0.422***	0.012***
				(564)	(559)	(0.044)	(0.004)	(0.008)	(0.000)
	Immigrant	0.081	0.106	2452	3344	0.449***	0.255***	0.379***	0.013***
				(569)	(544)	(0.077)	(0.008)	(0.015)	(0.001)
(6) Maternal education	Below 12	-0.477	-0.339	2339	3252	0.359***	0.266***	0.368***	0.012***
				(570)	(557)	(0.094)	(0.008)	(0.014)	(0.001)
	12-15	0.004	0.095	2430	3348	0.435***	0.291***	0.437***	0.013***
				(563)	(557)	(0.050)	(0.005)	(0.009)	(0.000)
	Above 15	0.663	0.677	2451	3417	0.527***	0.256***	0.381***	0.013***
				(562)	(529)	(0.079)	(0.010)	(0.020)	(0.001)
(7) Zip code median income	Bottom	-0.217	-0.139	2394	3285	0.386***	0.289***	0.407***	0.015***
				(568)	(565)	(0.076)	(0.007)	(0.013)	(0.001)
	Middle	0.121	0.085	2409	3337	0.450***	0.270***	0.407***	0.012***
				(568)	(554)	(0.072)	(0.007)	(0.014)	(0.001)
	Top	0.437	0.381	2436	3382	0.443***	0.264***	0.399***	0.011***
				(561)	(535)	(0.078)	(0.008)	(0.016)	(0.001)
(8) Maternal marital status	Non-married	-0.360	-0.234	2336	3237	0.363***	0.269***	0.384***	0.013***
				(573)	(565)	(0.076)	(0.006)	(0.011)	(0.001)
	Married	0.272	0.277	2459	3396	0.483***	0.292***	0.439***	0.012***
				(556)	(541)	(0.045)	(0.005)	(0.010)	(0.000)
(9) Maternal age at birth of children	Below 22	-0.396	-0.208	2268	3237	0.372***	0.268***	0.373***	0.011***
				(574)	(546)	(0.115)	(0.007)	(0.014)	(0.001)
	22-29	-0.006	0.075	2419	3357	0.442***	0.274***	0.415***	0.011***
				(561)	(543)	(0.059)	(0.006)	(0.011)	(0.001)
	30-35	0.278	0.305	2466	3390	0.484***	0.294***	0.446***	0.014***
				(557)	(559)	(0.069)	(0.007)	(0.015)	(0.001)
	Above 35	0.342	0.306	2480	3353	0.408***	0.326***	0.490***	0.018***
				(559)	(593)	(0.104)	(0.012)	(0.024)	(0.001)

Note: Descriptive statistics for each group are reported in columns (1) to (4). Columns (1) and (2) present mean combined mathematics and reading test scores for twins and singletons respectively. Columns (3) and (4) present mean and standard deviation of birth weight for twins and singletons respectively. Column (4) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (6) to (8) present estimates for singleton population. Column (6) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (7) 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 rage 847 to 3600 grams. Column (6) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. 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. Standard errors in column (5) are clustered at twin-pair level while in columns (6) to (8) at individual level. Sample sizes are: 126 826 individual-years observations in column (5), 5,757,145 individual-year observations in columns (6) and (8), 4,029,066 individual-year observations in column (7). There are fewer observations in zip code income because we do not observe these for years 1992 and 1993.

Table 4: Results by school quality measures

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
School quality measure	Sample	Mean test score		Mean (SD) birth weight		Pooled twin FE estimate	Singletons		
		Twins	Singletons	Twins	Singletons		Birth weight	Birth weight | gestation	Gestation
(1) Awarded grade	A	0.277	0.276	2437	3365	0.405***	0.274***	0.412***	0.012***
				(559)	(545)	(0.042)	(0.004)	(0.009)	(0.000)
	B	-0.095	-0.039	2410	3323	0.500***	0.284***	0.413***	0.012***
				(569)	(560)	(0.063)	(0.006)	(0.011)	(0.001)
	C & D & F	-0.399	-0.311	2375	3266	0.458***	0.275***	0.378***	0.014***
				(577)	(572)	(0.076)	(0.006)	(0.012)	(0.001)
(2) Average proficiency	Below median	-0.340	-0.249	2382	3281	0.438***	0.282***	0.395***	0.014***
				(580)	(568)	(0.061)	(0.005)	(0.010)	(0.000)
	Above median	0.297	0.298	2442	3371	0.423***	0.267***	0.404***	0.011***
				(555)	(544)	(0.043)	(0.004)	(0.009)	(0.000)
(3) Growth in proficiency	Below median	0.045	0.058	2420	3337	0.448***	0.286***	0.429***	0.012***
				(565)	(556)	(0.045)	(0.004)	(0.008)	(0.000)
	Above median	0.099	0.106	2422	3335	0.433***	0.285***	0.413***	0.013***
				(564)	(555)	(0.045)	(0.004)	(0.008)	(0.000)

Note: Descriptive statistics for each group are reported in columns (1) to (4). Columns (1) and (2) present mean combined mathematics and reading test scores for twins and singletons respectively. Columns (3) and (4) present mean and standard deviation of birth weight for twins and singletons respectively. Column (4) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (6) to (8) present estimates for singleton population. Column (6) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (7) 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 rage 847 to 3600 grams. Column (6) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. 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. Standard errors in column (5) are clustered at twin-pair level while in columns (6) to (8) at individual level. In the case of awarded grades since not all schools are awarded grades every year our sample consist of 124,380 observations used in models in column (5), 5,654,386 observations used in models in column (6) and (8) and 3,955,361 observations used in models in column (7). In the case of average proficiency and growth in proficiency we use 126,502 observations in models in column (5), 5,737,758 observations in models in columns (6) and (8) and 4,015,843 observations in models in column (7). 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).

Table 5: Results by school quality measures:
Running Florida data through other state school grading systems

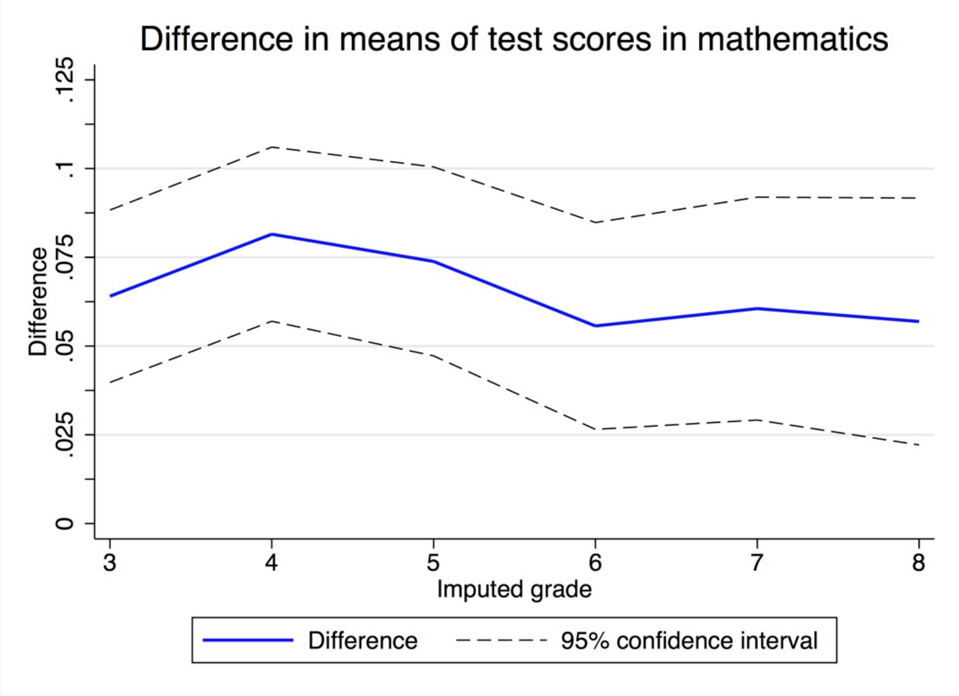
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Mean test score		Mean (SD) birth weight		Pooled twin FE estimate	Singletons		
State	Quality group	Twins	Singletons	Twins	Singletons		Birth weight	Birth weight | gestation	Gestation
(1) New York City	Top	0.305	0.291	2439	3363	0.388***	0.270***	0.406***	0.011***
				(560)	(546)	(0.049)	(0.005)	(0.010)	(0.000)
	Middle	0.052	0.073	2427	3336	0.490***	0.275***	0.407***	0.012***
				(563)	(557)	(0.051)	(0.005)	(0.009)	(0.000)
	Bottom	-0.181	-0.121	2395	3308	0.481***	0.294***	0.418***	0.014***
				(575)	(565)	(0.062)	(0.005)	(0.011)	(0.001)
(2) Louisiana	Top	0.373	0.371	2448	3381	0.395***	0.263***	0.403***	0.011***
				(556)	(541)	(0.048)	(0.005)	(0.010)	(0.000)
	Middle	-0.091	-0.024	2414	3325	0.481***	0.283***	0.409***	0.013***
				(568)	(560)	(0.054)	(0.005)	(0.010)	(0.000)
	Bottom	-0.490	-0.378	2365	3250	0.446***	0.267***	0.360***	0.015***
				(587)	(574)	(0.104)	(0.008)	(0.015)	(0.001)
(3) Indiana	Top	0.358	0.349	2448	3376	0.397***	0.260***	0.396***	0.011***
				(556)	(542)	(0.047)	(0.005)	(0.010)	(0.000)
	Middle	-0.068	-0.006	2412	3328	0.523***	0.286***	0.416***	0.013***
				(568)	(559)	(0.054)	(0.005)	(0.010)	(0.000)
	Bottom	-0.451	-0.353	2368	3259	0.430***	0.277***	0.384***	0.015***
				(586)	(574)	(0.097)	(0.007)	(0.014)	(0.001)

Note: Descriptive statistics for each group are reported in columns (1) to (4). Columns (1) and (2) present mean combined mathematics and reading test scores for twins and singletons respectively. Columns (3) and (4) present mean and standard deviation of birth weight for twins and singletons respectively. Column (4) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (6) to (8) present estimates for singleton population. Column (6) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (7) 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 rage 847 to 3600 grams. Column (6) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. 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. Standard errors in column (5) are clustered at twin-pair level while in columns (6) to (8) at individual 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 107950 observations used in models in column (5), 4976617 observations used in models in column (6) and (8) and 3474006 observations used in models in column (7). Louisiana system simulation consist of 109088 observations used in models in column (5), 5031341 observations used in models in column (6) and (8) and 3510705 observations used in models in column (7). Indiana system simulation consist of 107951 observations used in models in column (5), 4976769 observations used in models in column (6) and (8) and 3474098 observations used in models in column (7).

APPENDIX FOR ONLINE PUBLICATION ONLY

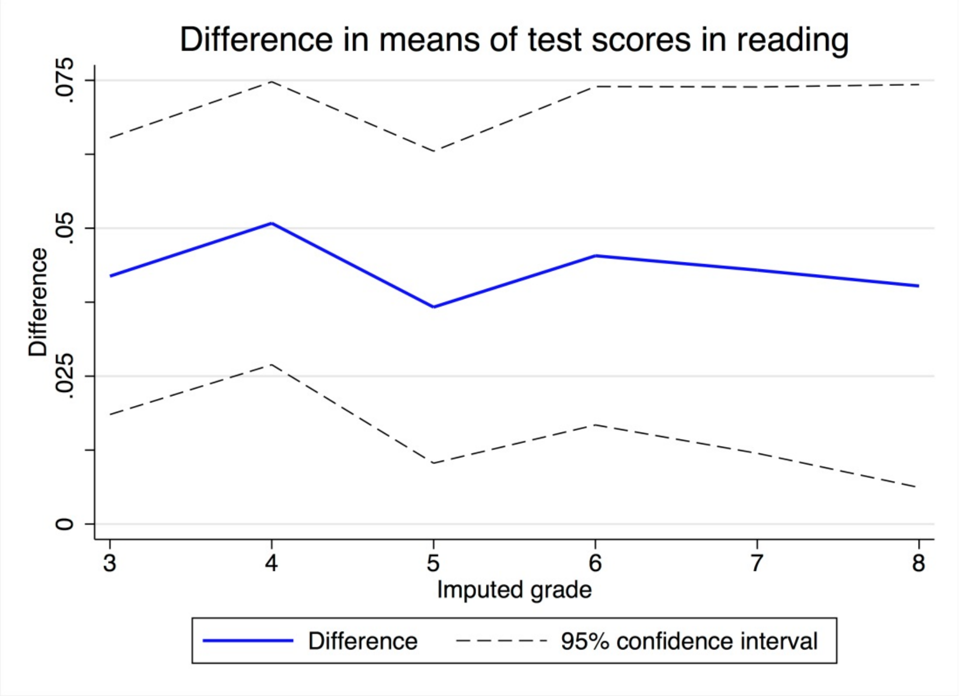
FIGURES

Figure A1. Average within-twin-pair difference in mathematics between heavier and lighter twins



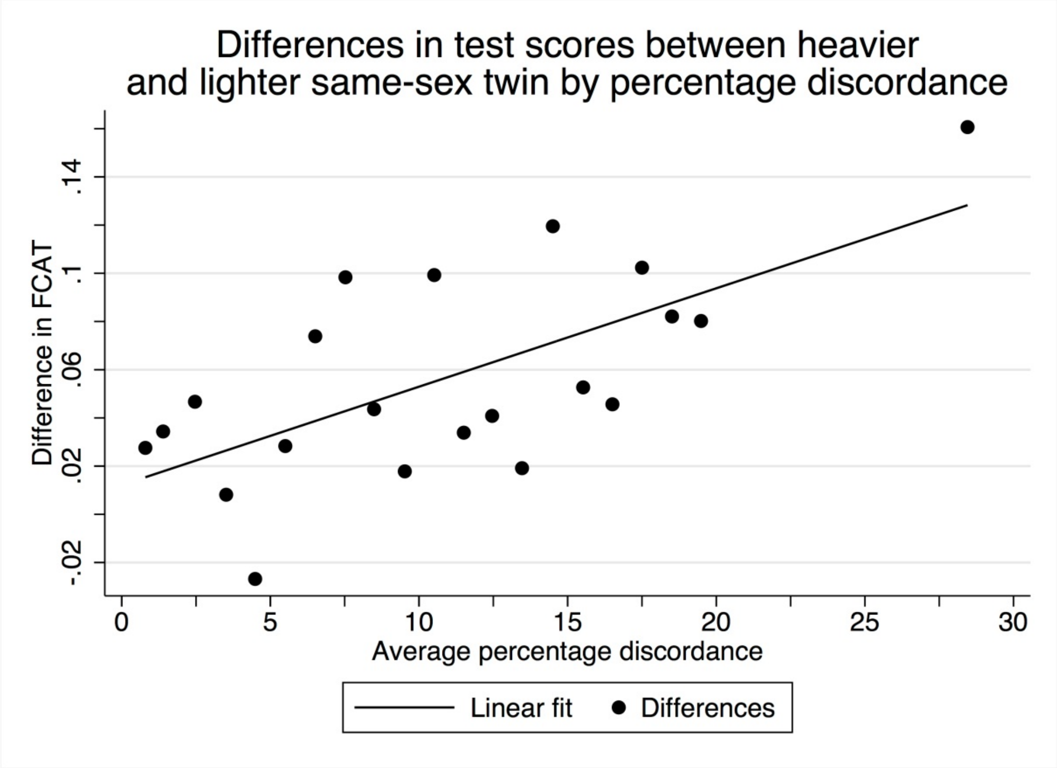
Note: Figure A1 plots difference between the test score in mathematics of heavier and lighter twin from each pair in each grade and the respective 95% confidence interval of this difference. In each grade we create an average of scores for heavier and lighter twins and then calculate the difference between the two.

Figure A2. Average within-twin-pair difference in reading between heavier and lighter twins



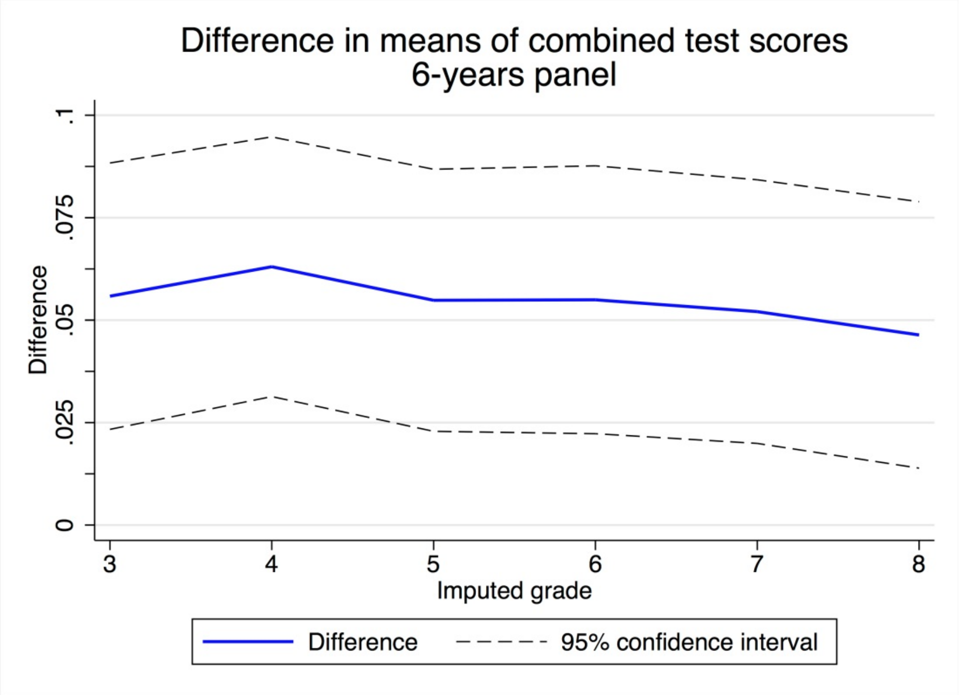
Note: Figure A2 plots difference between the test score in reading of heavier and lighter twin from each pair in each grade and the respective 95% confidence interval of this difference. In each grade we create an average of scores for heavier and lighter twins and then calculate the difference between the two.

Figure A3: Within twin-pair differences in test scores by discordance



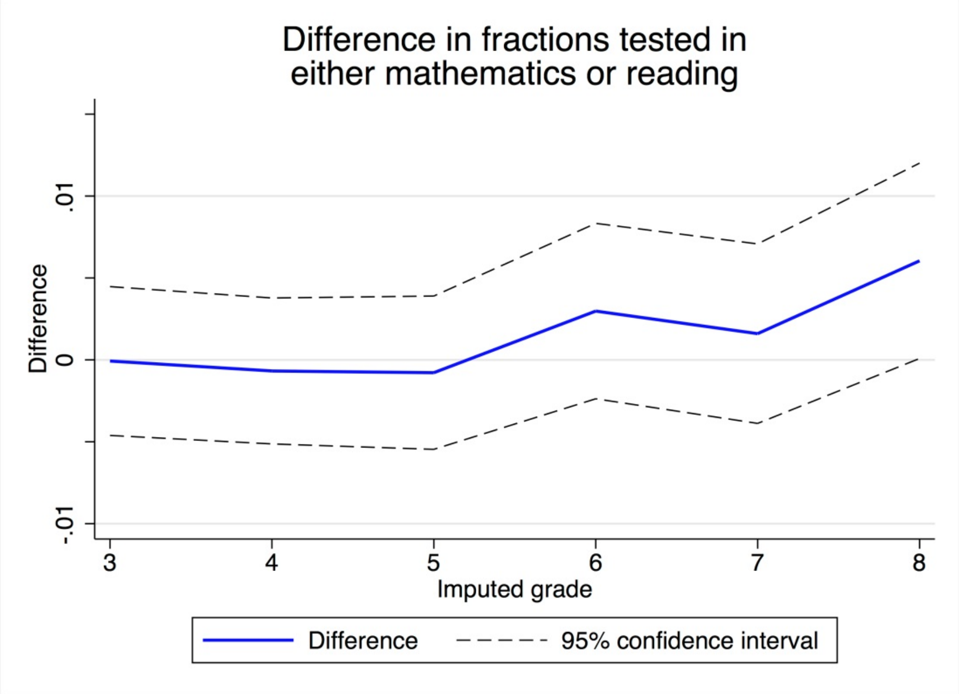
Note. Figure A3 plots mean differences in test scores between heavier and lighter twin for twin pairs of given discordance. Discordance levels are recorded every percent from 0 to 20 and then the last bin groups all twin pairs with discordance greater than 20 percent. Discordance is measured as difference in birth weight between heavier and lighter twin over the birth weight of heaver twin. Solid line fits a linear regression using the differences for each discordance bin as data.

Figure A4. Average within twin pair difference in test scores between the higher birth weight and the lower birth weight twin: Sample where both twins are observed in each of six grades



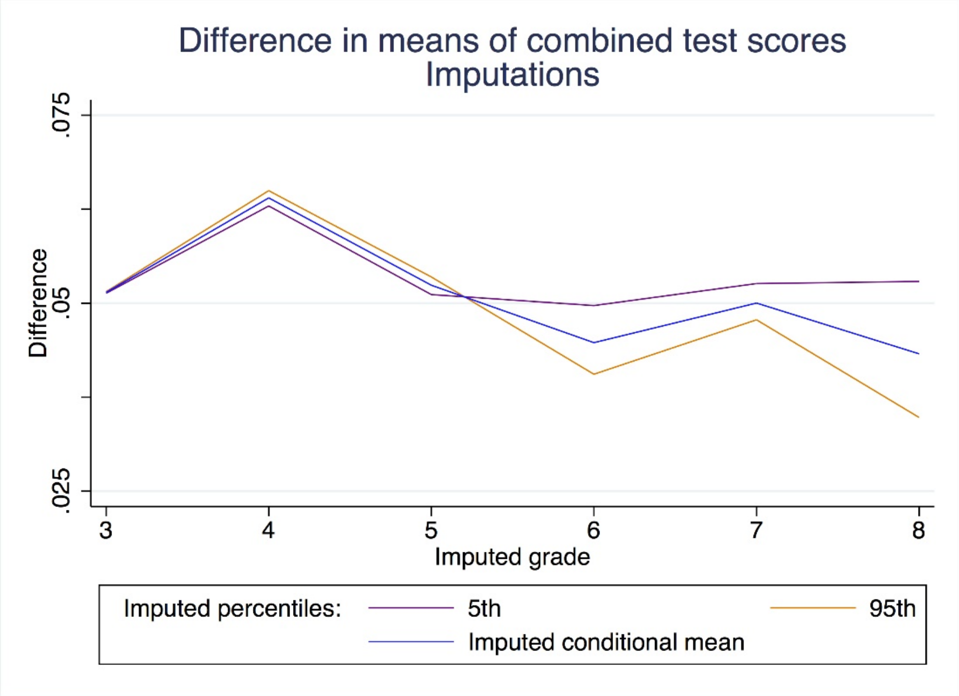
Note: Figure A4 plots the same difference as Figure 3 but for a 6-year panel of twin-pairs i.e., we restrict the sample only to individuals where we observe both twins mean test scores from grade 3 to grade 6.

Figure A5. Difference in fraction of lighter and heavier twins tested in either mathematics or reading



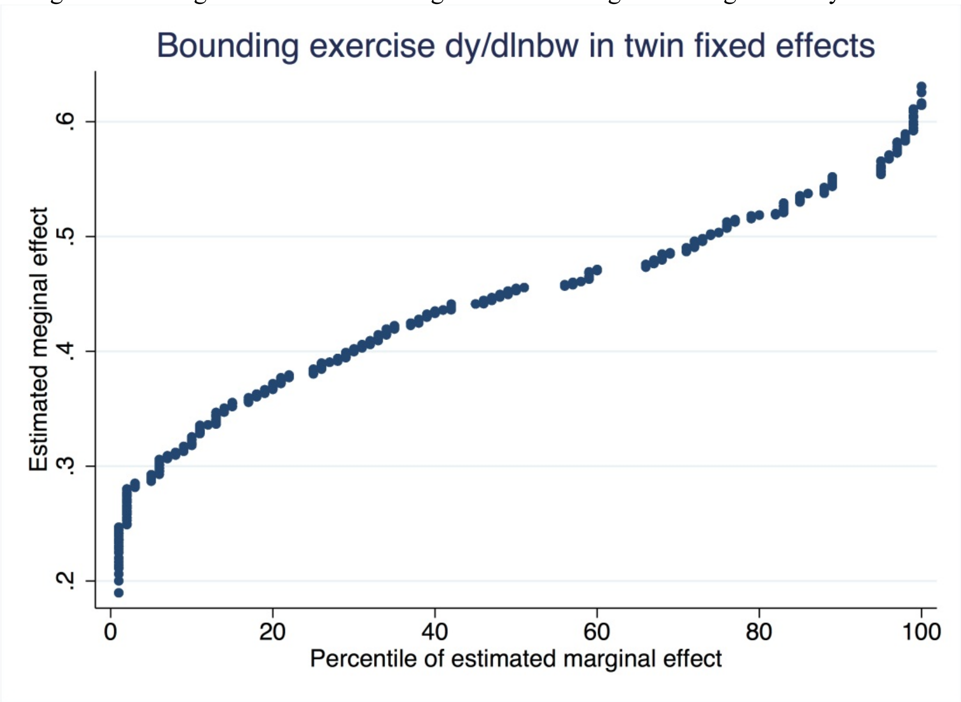
Note: Figure A5 plots difference and its 95% confidence interval of fraction of heavier and lighter twins attending each grade. We start with all twin pairs old for grade where at least one twin has been successfully matched to Florida public schools in 3rd grade. For each grade we then calculate the fraction of heavier and lighter individuals attending given grade and difference the two.

Figure A6: Differences across grades with 5th and 95th test score imputations for twins with missing scores



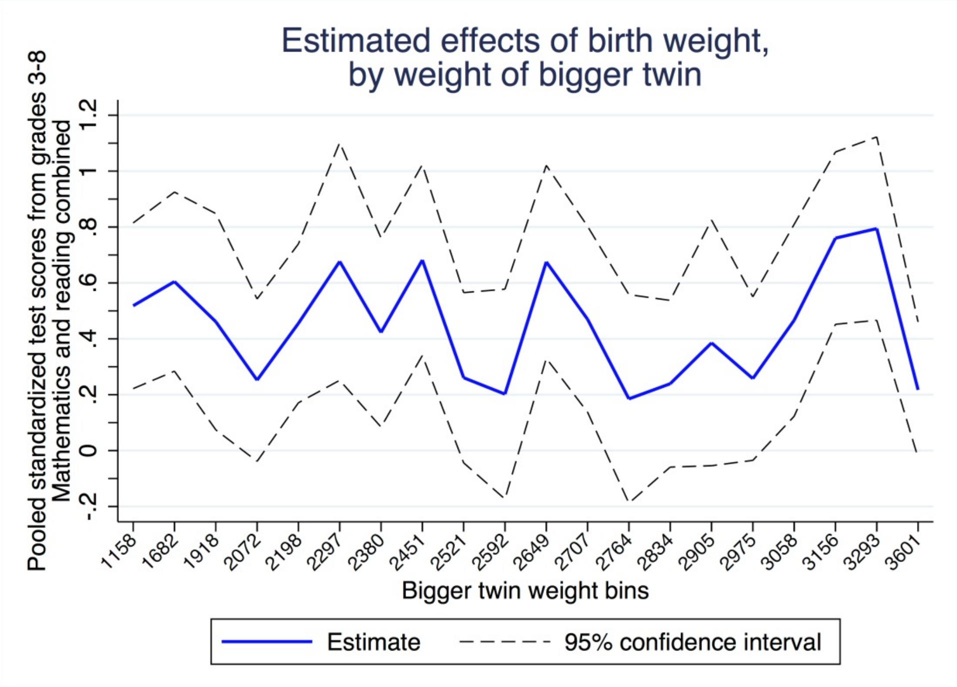
Note: Figure A6 plots three sets of differences calculated in the same way as in figure 3 but where we substitute the missing individual scores within twin pairs with either the 5th (solid purple line) or 95th (solid orange line) percentile or imputed conditional mean (solid blue line) of test scores in that grade.

Figure A7: Range of estimates of marginal effect of log birth weight in fully interacted model, twin FE



Note: Figure A7 plots the range of estimated marginal effect of log birth weight on test scores coming from a regression of test scores on log birth weight and its interactions with education, age and income categories as well as race, ethnicity, immigration and marital status dummies. This model is run in a twin fixed effects framework as in table 2. The percentile distribution of the estimated marginal effect of birth weight is on the X axis.

Figure A8. Estimated effects of birth weight, by weight of bigger twin



Note: Figure A8 plots coefficients and 95% confidence intervals from a twin FE regression where the dependent variable is the mean of pooled grades three to eight combined mathematics and reading test scores for each individual and the independent variables are 20 interactions corresponding to the product of log birth weight with indicators for 20 bins reflecting heavier twin percentiled birth weight. The regression additionally controls for infant gender and birth order within-twin pair. Heteroskedasticity robust standard errors are used to calculate the 95% confidence interval. Numbers on the x-axis correspond to the mean birth weight in each bin of heavier twin birth weight.

ONLINE APPENDIX TABLES

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

Sample		(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Pooled	Imputed grade					
			3	4	5	6	7	8
Total sample		0.441***	0.442***	0.524***	0.430***	0.426***	0.387***	0.372***
		(0.039)	(0.043)	(0.045)	(0.047)	(0.053)	(0.056)	(0.061)
(1) Gender	Boys	0.452***	0.473***	0.567***	0.386***	0.478***	0.413***	0.297***
		(0.068)	(0.073)	(0.079)	(0.080)	(0.094)	(0.096)	(0.101)
	Girls	0.445***	0.450***	0.489***	0.434***	0.453***	0.379***	0.428***
		(0.053)	(0.067)	(0.070)	(0.068)	(0.071)	(0.078)	(0.085)
(2) Maternal medical history	No medical problems	0.448***	0.460***	0.509***	0.454***	0.436***	0.382***	0.385***
		(0.048)	(0.055)	(0.056)	(0.060)	(0.063)	(0.073)	(0.077)
	Medical problems	0.415***	0.402***	0.530***	0.386***	0.394***	0.380***	0.331***
		(0.066)	(0.071)	(0.077)	(0.077)	(0.095)	(0.090)	(0.103)
(3) Maternal race (N=14 357)	White	0.464***	0.502***	0.541***	0.438***	0.417***	0.417***	0.388***
		(0.045)	(0.051)	(0.054)	(0.054)	(0.060)	(0.065)	(0.068)
	Black	0.387***	0.297***	0.481***	0.418***	0.452***	0.308***	0.346**
		(0.082)	(0.087)	(0.091)	(0.098)	(0.119)	(0.118)	(0.137)
(4) Maternal ethnicity	Non-Hispanic	0.433***	0.443***	0.516***	0.438***	0.395***	0.377***	0.356***
		(0.044)	(0.049)	(0.052)	(0.053)	(0.060)	(0.064)	(0.070)
	Hispanic	0.481***	0.445***	0.567***	0.399***	0.577***	0.436***	0.448***
		(0.079)	(0.094)	(0.093)	(0.103)	(0.110)	(0.115)	(0.125)
(5) Maternal immigration history	Non-immigrant	0.439***	0.470***	0.516***	0.438***	0.407***	0.366***	0.345***
		(0.044)	(0.049)	(0.052)	(0.053)	(0.061)	(0.065)	(0.070)
	Immigrant	0.449***	0.324***	0.561***	0.399***	0.511***	0.468***	0.476***
		(0.077)	(0.090)	(0.090)	(0.095)	(0.105)	(0.111)	(0.122)
(6) Maternal education	Below 12	0.359***	0.256**	0.483***	0.427***	0.253**	0.370**	0.343**
		(0.094)	(0.110)	(0.125)	(0.121)	(0.127)	(0.145)	(0.152)
	12-15	0.435***	0.466***	0.492***	0.412***	0.446***	0.366***	0.364***
		(0.050)	(0.055)	(0.055)	(0.060)	(0.070)	(0.070)	(0.078)
	Above 15	0.527***	0.517***	0.651***	0.486***	0.502***	0.483***	0.434***
		(0.079)	(0.089)	(0.099)	(0.097)	(0.099)	(0.123)	(0.130)
(7) Zip code median income (N=11 868)	Bottom	0.386***	0.426***	0.442***	0.300***	0.331***	0.407***	0.383***
		(0.076)	(0.083)	(0.085)	(0.090)	(0.110)	(0.110)	(0.133)
	Middle	0.450***	0.405***	0.529***	0.480***	0.496***	0.371***	0.325**
		(0.072)	(0.081)	(0.089)	(0.086)	(0.101)	(0.115)	(0.126)
	Top	0.443***	0.511***	0.539***	0.374***	0.374***	0.317***	0.422***
		(0.078)	(0.085)	(0.088)	(0.098)	(0.108)	(0.120)	(0.143)
(8) Maternal marital status (N=14 583)	Non-married	0.363***	0.339***	0.404***	0.417***	0.375***	0.364***	0.217*
		(0.076)	(0.083)	(0.086)	(0.090)	(0.112)	(0.113)	(0.115)
	Married	0.483***	0.497***	0.584***	0.443***	0.454***	0.401***	0.457***
		(0.045)	(0.050)	(0.053)	(0.055)	(0.058)	(0.064)	(0.072)
(9) Maternal age at birth of children	Below 22	0.372***	0.375***	0.409***	0.495***	0.236	0.396**	0.231
		(0.115)	(0.116)	(0.130)	(0.136)	(0.172)	(0.168)	(0.177)
	22-29	0.442***	0.417***	0.506***	0.375***	0.533***	0.416***	0.383***
		(0.059)	(0.067)	(0.066)	(0.071)	(0.082)	(0.085)	(0.093)
	30-35	0.484***	0.467***	0.583***	0.497***	0.468***	0.390***	0.426***
		(0.069)	(0.080)	(0.081)	(0.085)	(0.090)	(0.101)	(0.113)
	Above 35	0.408***	0.529***	0.563***	0.384***	0.172	0.264*	0.357**
		(0.104)	(0.114)	(0.135)	(0.125)	(0.135)	(0.156)	(0.155)

Note: Column (1) present pooled grade three through eight results for twin-FE model. Columns (3) to (8) present twin-FE estimates separately for each of the 6 grades. Models are the same as used in columns (2) and (3) to (8) in table 2. Sample size is 126 826 individual observations in pooled regressions in column (1) except for race, marital status and mean zip code income. In the case of race this discrepancy is caused by existence of other races with minor representation in Florida, in the case of income and marital status we do not have complete data for all mothers and residential locations. In all these cases the modified sample sizes are used. Each coefficient comes from a separate regression.

Table A2: Sensitivity of results to model specification

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sample	Pooled	Imputed grade					
		3	4	5	6	7	8
(1)	0.441***	0.442***	0.524***	0.430***	0.426***	0.387***	0.372***
ln(birth weight)	(0.039)	(0.043)	(0.045)	(0.047)	(0.053)	(0.056)	(0.061)
(2)	0.529***	0.581***	0.662***	0.467***	0.524***	0.409***	0.424***
Both twins above 2500g	(0.098)	(0.110)	(0.115)	(0.121)	(0.126)	(0.144)	(0.152)
(3)	0.473***	0.471***	0.569***	0.497***	0.456***	0.339***	0.418***
Both twins below 2500g	(0.061)	(0.066)	(0.070)	(0.074)	(0.087)	(0.089)	(0.098)
(4)	0.518***	0.389***	0.536***	0.572***	0.590***	0.503***	0.581***
Both twins 1500g-2499g	(0.082)	(0.092)	(0.097)	(0.105)	(0.114)	(0.120)	(0.131)
(5)	0.589***	0.609***	0.724***	0.747***	0.517**	0.393*	0.309
Both twins <1500g	(0.152)	(0.177)	(0.158)	(0.196)	(0.209)	(0.229)	(0.272)
(6)	0.186***	0.185***	0.222***	0.178***	0.180***	0.170***	0.155***
Birth weight in 1000g	(0.017)	(0.019)	(0.019)	(0.020)	(0.023)	(0.024)	(0.026)
(7)	0.197***	0.196***	0.233***	0.190***	0.193***	0.177***	0.171***
Birth weight	(0.017)	(0.019)	(0.020)	(0.021)	(0.023)	(0.025)	(0.027)
Birth weight x (birth weight -	-0.106***	-0.117***	-0.106***	-0.117***	-0.108**	-0.059	-0.114**
mean twin pair birth weight)	(0.032)	(0.036)	(0.037)	(0.039)	(0.045)	(0.047)	(0.051)

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

Table A3: Results by school quality measures and predicted SES

Predicted SES	School quality	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Mean test score		Mean (SD) birth weight		Pooled twin FE estimate	Singletons		
		Twins	Singletons	Twins	Singletons		Birth weight	Birth weight | gestation	Gestation
(1) Bottom	A	-0.294	-0.186	2333	3228	0.298***	0.263***	0.359***	0.013***
				(566)	(560)	(0.088)	(0.008)	(0.016)	(0.001)
	B	-0.439	-0.343	2335	3216	0.439***	0.257***	0.343***	0.013***
				(570)	(561)	(0.098)	(0.008)	(0.017)	(0.001)
	C & D & F	-0.610	-0.502	2324	3193	0.418***	0.259***	0.332***	0.015***
				(580)	(569)	(0.096)	(0.008)	(0.016)	(0.001)
(2) Middle	A	0.201	0.211	2454	3375	0.406***	0.273***	0.425***	0.011***
				(557)	(537)	(0.068)	(0.007)	(0.014)	(0.001)
	B	0.005	0.021	2451	3366	0.569***	0.282***	0.443***	0.010***
				(558)	(548)	(0.101)	(0.009)	(0.018)	(0.001)
	C & D & F	-0.127	-0.130	2450	3351	0.632***	0.269***	0.408***	0.011***
				(557)	(558)	(0.124)	(0.011)	(0.022)	(0.001)
(3) Top	A	0.633	0.580	2476	3432	0.455***	0.263***	0.411***	0.011***
				(551)	(530)	(0.067)	(0.007)	(0.014)	(0.001)
	B	0.363	0.346	2479	3428	0.465***	0.303***	0.438***	0.014***
				(569)	(547)	(0.135)	(0.012)	(0.024)	(0.001)
	C & D & F	0.178	0.177	2499	3418	0.277	0.315***	0.466***	0.016***
				(568)	(561)	(0.250)	(0.017)	(0.034)	(0.002)

Note: Descriptive statistics for each group for the whole populations are reported in columns (1) to (4). Columns (1) and (2) present mean combined mathematics and reading test scores for twins and singletons respectively. Columns (3) and (4) present mean and standard deviation of birth weight for twins and singletons respectively. Column (4) presents pooled grades three through eight twin-FE model estimates corresponding to model outlined in column (2) in table 2. Columns (6) to (8) present estimates for singleton population. Column (6) presents the correlation between pooled grades three through eight test scores and birth weight for all singletons. Column (7) 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 rage 847 to 3600 grams. Column (8) presents the correlation between pooled grades three through eight test scores and gestation weeks for all singletons. 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. Standard errors in column (5) are clustered at twin-pair level while in columns (6) to (8) at individual level. Sample sizes in column (5) are 17172, 8913 and 14744 for each row in panel 1 and 28402, 7739 and 5711 for each row in panel 2 and 33673, 5022 and 2676 for each row in panel 3, respectively. Sample sizes in columns (6) and (8) are 774228, 409546 and 703874 for each row in panel 1 and 1 202773, 386498 and 329796 for each row in panel 2 and 1430736, 263097 and 153939 for each row in panel 3, respectively. Sample sizes in column (7) are 520564, 284113 and 504120 for each row in panel 1 and 800480, 272933 and 260850 for each row in panel 2 and 1003836, 192600 and 115865 for each row in panel 3, respectively.