Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-17-16

School Starting Age and Cognitive Development

Elizabeth Dhuey
Associate Professor of Economics
University of Toronto

David Figlio
Orrington Lunt Professor of Education and Social Policy and of Economics
IPR Fellow
Northwestern University

Krzysztof Karbownik
IPR Research Associate
Northwestern University

Jeffrey Roth
Research Professor of Pediatrics
University of Florida

Version: August 24, 2017

DRAFT

Please do not quote or distribute without permission.

ABSTRACT

The researchers present evidence of a positive relationship between school starting age and children's cognitive development from age 6 to 15 using a regression discontinuity design and large-scale population-level birth and school data from the state of Florida. They estimate effects of being relatively old for grade (being born in September versus August) that are remarkably stable –always just around 0.2 SD difference in test scores – across a wide range of heterogeneous groups, based on maternal education, poverty at birth, race/ethnicity, birth weight, gestational age, and school quality. While the September-August difference in kindergarten readiness is dramatically different by subgroup, by the time students take their first exams, the heterogeneity in estimated effects effectively disappears. They document substantial variation in compensatory behaviors targeted towards young for grade children. While the more affluent families tend to redshirt their children, young for grade children from less affluent families are more likely to be retained in grades prior to testing. School district practices regarding retention and redshirting are correlated with improved outcomes for the groups less likely to use those remediation approaches (i.e., retention in the case of more-affluent families and redshirting in the case of less-affluent families.) They also study college and juvenile detention outcomes using administrative data from a large Florida school district, and show that being an older age at school entry increases children's college attainment and reduces the likelihood of being incarcerated for juvenile crime.

1 Introduction

Parents often start thinking about enrolling their children in primary school during infancy and toddlerhood. There are logistics questions like transportation and school supplies, but another question that they now have to consider is whether their child is old enough to start schooling. This question has been made more fraught as the popular media has begun to report on research findings around effects of age at school entry (Weil 2007). In response to these concerns, an increasing number of parents in the United States have been delaying sending their children to kindergarten because they believe doing so will give them an advantage over their peers, whether academically, socially, or even athletically (Deming and Dynarski 2008). This practice is called redshirting: the delaying a child's school entry until they can develop further, much like a college sports team might keep an athlete out of varsity competition for a year while they develop additional skills. In fact, Bassok and Reardon (2013) find that between 4 and 5.5 percent of children delay entry into kindergarten in the United States.

Most states in the United States and many jurisdictions worldwide have a single specific cutoff date which determines when a student can enter primary school. For example, in Florida, a child is eligible to enter kindergarten if s/he turns five years old by September 1st of the relevant school year. These cut off dates create a distribution of ages of children within each cohort at school entry, where the oldest child will be approximately one year older than the youngest child. A number of recent studies have found that children who are relatively older than their classmates at the beginning of primary school have a variety of short- and medium-run advantages such as scoring higher on standardized exams through primary and secondary school, having higher development of non-cognitive skills, and being less likely to commit a crime as a teenager or an adult.[1] These findings suggest that early differences in maturity can propagate through the human capital accumulation process into later life and may have important implications for adult outcomes and productivity. At the same time, the relationship between being relatively older at school entry and a variety of adult outcomes is more mixed, and previous research has generated inconclusive results.[2]

It can be challenging to estimate the effects of school starting age, because a student's age when entering primary education can be manipulated (via birth timing and/or redshirting) and may be correlated with family background characteristics. It has been shown that seasonal birth rates (which affects age relative to a cutoff) may vary based on family background characteristics (Buckles and Hungerman 2013). Research has also shown seasonal patterns in birth outcomes, mental health, neurological disorders, adult height, life expectancy, intelligence, and income (Currie and Schwandt 2013). There is evidence that conditions at conception, such as in utero exposure to illness/disease (Currie and Schwandt 2013) or nutrient deprivations due to seasonal nutritional intake (Barker 1990), may have an effect as well. Relatedly, we also know that parents can manipulate when children start school by redshirting. These redshirted children tend to be more likely male, white and from high socioeconomic statuses (Bassok and Reardon 2013). Dobkin and Ferreira (2010) find similar attributes of redshirted students using data from California and Texas – higher education levels, higher incomes, and lower likelihood of being African-American or Hispanic. As

[1]See for example, Bedard and Dhuey (2006), Datar (2006), Crawford et al. (2007), Puhani and Weber (2007), McEwan and Shapiro (2008), Elder and Lubotsky (2009), Smith (2009), Kawaguchi (2011), Cook and Kang (2016), Lubotsky and Kaestner (2016), McAdams (2016), Landerso et al. (2017b) and Attar and Cohen-Zada (2017). Some other outcomes investigated in this literature include high school leadership (Dhuey and Lipscomb 2008), pre-academic university tracks and enrollment in selective universities (Bedard and Dhuey 2006), and disability identification and special education (Dhuey and Lipscomb 2010; Elder and Lubotsky 2009; Evans et al. 2010; Morrow et al. 2012)

[2]For instance, Fredriksson and Ockert (2014) and Kawaguchi (2011) find that older children at school entry earn higher wages; and Du et al. (2012) find that they are more likely to be a corporate CEO. In contrast, Black et al. (2011), Dobkin and Ferreira (2010), and Fertig and Kluve (2005) find no such long-term wage effects.

a consequence, comparing children based on their age when starting school is often fraught with omitted-variables concerns, and even results from studies with sufficient numbers of observations to make use of regression discontinuity evidence – say, comparing September births to August births in locales with a September 1st cutoff for school entry – may still be subject to omitted-variables bias due to endogenous birth timing.

We use detailed population-level administrative data from the state of Florida, where we observe matched birth and schooling outcomes, to study this question. In doing so, we make four principal contributions to the literature on the effects of school starting age. First, we offer the most comprehensive set of controls for potential selection into timing of birth yet considered in the literature, and bring together in the same research design the two most compelling approaches used in the literature to attempt to correct for this selection. Specifically, we present the first evidence from an environment in which we can execute a regression-discontinuity design, comparing children whose ages mean that they would "naturally" be the oldest in their class to those whose ages mean that they would "naturally" be the youngest in their class, while at the same time making this comparison *within families*. Comparing one child born in August to their sibling born in September dramatically reduces the likelihood that observed results are due to unobserved differences in families who time births for August versus those who time births for September.[3] Some studies (e.g., Cook and Kang (2016) and Elder and Lubotsky (2009)) have made use of the regression discontinuity approach before, and one study (Black et al. 2011) has made sibling comparisons, but we are the first to simultaneously compare siblings who just barely met or missed the threshold for school attendance in a given academic year.[4] We also are able to control for conditions and treatments surrounding pregnancy and birth. We ultimately find that these extra controls do not alter our results, indicating that omitted-variables bias in the extant literature is likely not as large as some might fear *ex ante*.

Given this evidence, we make a second contribution by presenting US population-level evidence on downstream outcomes such as college attendance and completion and juvenile incarceration. We use data from a large Florida county where we lack sufficient observations to make a direct within-family comparison but can carry out the regression discontinuity design to study these important later life outcomes. Like Cook and Kang (2016), we find evidence that August-born children are more likely than September-born peers to be incarcerated in their teenage years. (Cook and Kang (2016) study juvenile delinquency but not the rarer juvenile incarceration.) We also make use of data matching school records to administrative college attendance records from the National Student Clearinghouse and show that August-born children are less likely to attend or complete college, or to complete college at a selective institution.[5]

Our third contribution involves a comprehensive study of the heterogeneous effects of school starting age. Families differ dramatically in terms of the degree to which they actively attempt to remediate their children's being relatively young for grade. Schanzenbach and Howard (2017), for example, report that summer-born sons of college-educated parents are nearly four times as likely to be redshirted as are summer-born sons of high-school educated parents. If families differ this remarkably regarding how they treat relatively young children, it stands to reason that the effects

[3]Of course, it's always possible that a family might, for some reason, intentionally time a birth for September but not do so for another birth, but at least any characteristic of a family that is invariant across siblings will be absorbed in the family fixed effect.

[4]In this study, we are unable to decompose our "relative age" effect into the three separate components: effect of a child's age at school entry, effect of their age at the time of outcome measurements, and the effect of their age relative to their peer group (Cascio and Schanzenbach 2016).

[5]Dobkin and Ferreira (2010) study a variant of this question using a regression discontinuity design and Census data, where college attainment is self-reported and where it is impossible to identify specific forms of postsecondary education or educational selectivity. They limit college attainment to "some college or higher" or "associate degree or higher", but do not find evidence of a difference in college attendance or completion.

of school starting age might be different for different groups of families. To date, however, there has been little research showing the heterogeneous effects of school starting age in the US context, largely due to limitations in US administrative data.[6] Elder and Lubotsky (2009) estimate different effects of school starting age by quartile of family background, but because they use survey data they do not have sufficient observations to carry out this heterogeneity analysis using the preferred regression discontinuity approach. Cook and Kang (2016) use population-level data and a regression discontinuity analysis, but because they focus on crime and delinquency they only investigate various definitions of significant disadvantage. To the best of our knowledge, this paper represents the first instance of an analysis of heterogeneous impacts of school starting age – both for high-SES and low-SES families – in a regression discontinuity framework using US data. Moreover, we consider a wide range of cuts of the data. We stratify by maternal education; by poverty at birth; by race and ethnicity; by birth weight; by gestational age; and by experienced school quality; as well as by gender interacted with many of these stratifications. These stratifications are potentially important because they illustrate how age effects might differ depending on generalized school factors or by biological factors. For example, we know that better neonatal health, as proxied by higher birthweight, has a positive effect on longer-run outcomes such as educational attainment, IQ, and life-cycle earnings (Black et al. 2007; Figlio et al. 2014; Bharadwaj et al. 2017). Therefore, it is natural to think that maybe birthweight might interact with a child's age relative to their classmates within the human capital production function framework (Cunha et al. 2010). This interaction could also occur to the degree to which educators have difficulty distinguishing between innate ability and maturity. Birthweight and its subsequent effect on childhood height and weight may make it difficult to disentangle maturity from ability as larger children may appear to be more mature due to their physical stature. Likewise, gestational age is another avenue one might suspect could affect the age gap (see Figlio et al. (2016) and Garfield et al. (2017) for the analysis of the role of gestational age in cognitive development). These interactions between initial birth endowments and school starting age has never been studied in the extant literature.

We find remarkable stability in the effects of school starting age across exceptionally different groups of people – despite differences in remediation strategies. We also find that the August-September difference in test scores is present and approximately equal in high-quality and lower-quality schools alike. This pattern of results suggests that the remediation for being relatively young for grade may be more challenging than those who seek to remediate might believe. This finding of an exceptional lack of heterogeneous effects of school starting age leads us to our fourth contribution. In this paper we directly explore the potential efficacy of attempted remediation techniques. Like Schanzenbach and Howard (2017), we show in our population-level data that there exist substantial differences in remediating behaviors among parents of different socioeconomic groups, with higher-SES parents being more likely to redshirt their children than lower-SES parents. Conversely, children who are from lower-SES families are more likely than their higher-SES counterparts to be retained in an early grades. As a potential consequence of these two sets of actions, by the time children reach third grade, the ratio of September to August-born children who are below grade for age is roughly equal across SES groups. This pattern of behaviors could help to explain why we document such a

[6] School registers in the US rarely contain background variables other than race/ethnicity and free lunch status, so only with either a match to birth certificates or the use of Census style data sets can researchers study heterogeneous effects with regard to a wide range of background factors. Researchers have studied heterogeneity in settings with broader access to registry data, and thus background variables: Chile (McEwan and Shapiro (2008), who find little differences in the effects of SSA by parental education); Denmark (Landerso et al. (2017b), who find evidence for smaller adverse effects of SSA on crime for groups with both better educated mothers and unemployed fathers); Israel (Attar and Cohen-Zada (2017), who find little differences by parental education; Norway (Black et al. (2011), who find little differences by predicted family affluence); and Sweden (Fredriksson and Ockert (2014), who find larger advantage in both education and earnings for children of lower educated parents).

strong SES gradient in the September-August difference in kindergarten readiness (where high-SES families are disproportionately likely to redshirt August-born children) but no SES gradient in the September-August difference in third grade test scores.

Armed with this evidence, we then turn to the question: Are remediation approaches like redshirting or grade retention more effective when used by groups for whom the approach is unusual? While we cannot obtain strong causal evidence on this point, we produce suggestive evidence that indicates that this may be the case. Florida has large county-level school districts that vary dramatically in the rate of redshirting or retention of August-born children. Medium-to-large Florida school districts range in their August-born redshirting rates from fewer than two percent to over ten percent, and range in their August-born early-grade retention rates from 20 percent to 45 percent. Districts with relatively high redshirting rates have higher-than-usual redshirting rates for both low-SES and high-SES August-born children alike (the correlation between overall August redshirting rates and low-SES August redshirting rates in these districts is 0.737) and districts with relatively high early-grade retention rates have higher-than-usual early-grade retention rates for all SES groups (the correlation between overall August early-grade retention rates and high-SES August early-grade retention rates in these districts is 0.745). We find that districts where redshirting is more prevalent have lower August-September differences in test scores for low-SES families (for whom redshirting is less common), and that districts where early-grade retention is more prevalent have lower August-September differences in test scores for high-SES families (for whom early-grade retention is less common). These findings, while merely suggestive, indicate a potential role for strategically-deployed instructional policies and practices to help modifying preparation differences caused by school starting age cutoffs.

2 Estimation

2.1 Data

We used birth records from the Florida Department of Health for all children born in Florida between 1994 and 2000, merged with school records maintained by Florida Department of Education for the academic years 1997-98 through 2011-12. The children were matched along four dimensions: first and last names, date of birth, and social security number. Rather than conducting probabilistic matching, the match was performed such that a child would be considered matched so long as (1) there were no more than two instances of modest inconsistencies, 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 confirmed a very high and clean match rate. In the overall match on the entire 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 likely due to typos in the birth or school records.

There were 1,220,803 singleton births with complete demographic information in Florida between 1994 and 2000, and of these 989,054 children were subsequently observed in Florida public schools data, representing an 81.0 percent match rate.[7] The match rate is almost identical to the percentage of children who are born in Florida, reside there until schooling age, and attend public school, as computed using data from the decennial Census and American Community Survey for years 2000 through 2009 (Figlio et al. 2014). Siblings are identified in school districts representing the vast majority of Florida households; Figlio et al. (2014) discuss the differences between these school districts, which are disproportionately nonrural, and the state as a whole.

[7] We exclude multiple births from the analysis.

We have a wide variety of demographic characteristics of the mother that are gathered from the Florida birth certificate. These include racial-ethnic information, education level, marital status at the time of the child's birth, and place of residence. We also have demographic characteristics of the father if he appears on the birth certificate, and health and demographic characteristics of the newborn. We observe birth weight, gestational age and indicators for any maternal health problems, whether or not they are related to the pregnancy. Finally, we know if the birth was paid for by Medicaid, an indicator of living in or near poverty at the time of the birth.

Table A1 documents demographic differences between the full population of births and the set of families whom we include in the empirical analysis. First, it is worth noting that August and September births do not appear to differ substantively from all Florida births (columns 1 and 2) suggesting that seasonality in birth characteristics might be less of a problem in this analysis as compared to some other studies. That said, these averages may still mask substantial heterogeneity. Comparing columns 2 and 3 reveals the cost of only being able to utilize students attending public schools and remaining in these schools until at least third grade, where we first observe their test scores, as the sample used in the analysis is negatively selected compared to full population of births. In that children observed in public schools are more likely to be African-American (25.8% vs. 22.4 %), less likely to have college educated (15.2% vs. 20.1%) or married (60.7% vs. 65.2%) mother and more likely to utilize Medicaid payments during birth (50.8% vs. 45.1%). Most of these differences are due to the fact that more affluent families are more likely to send their children to private schools than do less affluent families, rather than any substantial additional selection occurring between school start and third grade.

More to the point of the present paper, it is also the case that fewer September-born children than August-born children are enrolled in public school at least through third grade. If the "missing" September children have particularly favorable or unfavorable academic achievement potential it could bias our school starting age estimates. The August-September gap in demographic characteristics among the full population and children included in the analysis is similar across most dimensions except for maternal education and poverty. September-born children who are not included in the analysis are more likely to have high school educated mothers and less likely to have college educated mothers, and are more likely to have Medicaid-funded births, relative to their August counterparts, and also in comparison to the same difference among children whom we include in the regression sample. However, these differences are small and never exceed five percent of the mean value for a given characteristic.[8] That said, in Section 3 below we formally document these potential selection issues and carry out a bounds analysis to determine the degree to which they might influence our results.

In our data we can also observe school quality as defined by the state of Florida via its school accountability system. Since 1999, the Florida Department of Education has awarded each of its public schools a letter grade ranging from A (best) to F (worst). Initially, the grading system was based mainly on average proficiency rates on the FCAT standardized exam. 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; other quality indicators, such as competency in science, were subsequently added. We utilize this information to construct a contemporaneous school quality

[8] In addition, comparing columns 3 and 6 in Table A1 demonstrates that the sample of siblings observed in Florida schools is modestly positively selected as compared to all students born in August or September and attending public schools. Children with siblings in our sample are more likely to have mothers who are college educated (19.9% vs. 15.2%) and married (63.4% vs. 60.7%) at the time of birth. However, the August-September *differences* in observable characteristics that have the potential to undermine our regression discontinuity design are relatively small in both singleton and sibling samples which alleviates concerns about endogenous timing of birth by parents with particular characteristics. In fact, for neither of the variables is the difference larger than 2.6 percent.

measure. For each school, we compute a simple average of the observed gain scores between 2002 and 2013, as measured by the Florida Department of Education, which we then convert into a percentile rank in the observed gains distribution across Florida schools. These values are then attached to students for each school year and school they attend.

In this paper we focus on four short- or medium- term outcomes in our statewide analysis: kindergarten readiness, parental holding back behavior (redshirting), school retention behavior and test scores. Kindergarten readiness is measured by a universally-administered screening at the entrance to kindergarten. The Florida Department of Education recorded readiness measures for those who entered kindergarten in fall 2001 and before, and those who entered kindergarten in fall 2006 or later.[9] Because of this data restriction we are unable to use this outcome for children born between 1997 and 1999. Holding back or redshirting is defined as an indicator variable that equals to one if a child has higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one.[10] These are six or above for kindergarten and seven or above for grade one. We view this variable as primarily a parental decision. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade.[11] Finally, our measure of academic performance is based on Florida Comprehensive Assessment Test (FCAT) in mathematics and reading, a state-wide standardized yearly assessment of all students in Florida conducted in grades three through ten. In this paper we focus on test scores in grades three through eight, because curriculum differences make interpersonal test score comparisons relatively difficult in high school (e.g., one tenth grader is taking algebra while another is enrolled in calculus). Therefore, each child in the sample can contribute up to six observations, one for each grade observed. For brevity we average the math and reading test scores but we present main results split by reading and math in the Appendix Table A2. These are qualitatively similar across the two tests with quantitatively larger estimates for reading.

We supplement the main results with longer-run analysis of college outcomes and juvenile incarceration before age 16, based on a separate data set for the 1981 to 1990 birth cohorts from a large, diverse and unnamed county in Florida. To facilitate this analysis, student records have been merged to National Student Clearinghouse, providing us with information on college attendance, graduation and institution selectivity. We define a competitive/selective colleges as those rated "competitive" or above as recorded in the NCES-Barron's Admissions Competitive Index Data Files for 2004.[12] The school district also recorded which students were incarcerated in the juvenile justice system before their 16th birthday. In this school district, unlike in our primary statewide analysis sample, we cannot measure variables observed on birth certificates, but we are able to measure whether a child was limited English proficient, received subsidized meals in school, or lived in a low-income neighborhood.[13] The final September vs. August singleton estimation sample for college outcomes and

[9] In the early round of kindergarten readiness assessments, teachers administered a readiness checklist of academic and behavioral skills designed by the state Department of Education with a dichotomous ready/not-ready measure recorded in state records. In the later round of kindergarten readiness, the state universally implemented the DIBELS assessment aimed at measuring early pre-literacy skills. DIBELS uses a discrete measure that we dichotomize using the approach described in Figlio et al. (2013) so that the percent identified as kindergarten ready corresponds to the percentage in the later assessment. In our analysis sample, the birth cohorts between 1994 and 2000 who took the kindergarten readiness assessment, therefore, are those born between 1994 and 1996 (kindergarten checklist) and those born in 2000 (DIBELS).

[10] Kindergarten attendance in Florida is not mandatory but it is heavily subsidized and 95.8% of children in school records whom we observe in grade one also attended kindergarten. In our estimation sample this fraction is 89.9%.

[11] Florida has mandatory retention policy in grade three, and thus we are unable to credibly utilize retention as school behavior measure after grade two (Schwerdt et al. 2015; Ozek 2015).

[12] We have also restricted the analysis to those institutions rated "very competitive" or above and the results, available upon request, are highly similar.

[13] We observe neighborhood affluence at a very fine level, as the school district divides county into over 1,000 micro

juvenile incarceration consists of 16,923 individuals; however, we cannot perform a further sibling comparison in the regression discontinuity analysis in this county due to lack of statistical power.

2.2 Methods

We begin with a simple model of the relationship between student outcomes and month of birth. In the main specification we restrict our attention to the August-September comparison, where September born children are about one year older than August born children at the time of school entry. We estimate the following equation:

$$Y_i = \beta Sept_i + \gamma X_i + \varepsilon_i \tag{1}$$

where Y_i is one of the outcome variables for child i as defined in Section 2.1: kindergarten readiness; test scores in grades 3 to 8; being redshirted; being retained in an early grade; as well as our juvenile incarceration and post-secondary outcome measures. $Sept_i$ is an indicator variable for being born in the month of September; X_i contains mother and child control variables including year of birth dummies, log birth weight, gestational length, start of prenatal care in first trimester, and indicators for congenital anomalies, abnormal conditions at birth and maternal health problems; ε_i is the error term.[14] In order to maintain as balanced sample as possible we estimate redshirting and retention behaviors for the population where we also observe test scores.[15]

In Equation 1 we do not include any demographic controls since we also present heterogeneity analyses utilizing these covariates. However, we do control for birth endowments of children as they may vary within a year. These include log birth weight, gestational age, and indicators for start of prenatal care in first trimester, congenital anomalies (e.g. spina bifida, Down syndrome), abnormal conditions at birth (e.g. meconium aspiration syndrome, seizures) and maternal health problems (e.g. hypertension, diabetes). The parameter of interest, β, is the causal impact of age under the assumption that the unobservables are not correlated with month of birth. The exogenous variation in school starting age comes from variation in month of birth (August vs. September) and the administrative school starting rule in Florida (September 1st), thus generating a regression discontinuity design. The identifying assumption can be then translated into the following statement: children born in August and September are identical on observable and unobservable characteristics except for the age at which they begin schooling. It has been documented, however, that births exhibit seasonality and parental observable characteristics differ over the course of the year. In the case of Florida we also find that being born in September is correlated with observable family characteristics e.g. better educated and Hispanic mothers are less likely to have September births while mothers with Medicaid births are more likely to deliver in September. These differences are generally small – effect sizes between 0.2 percent for the African-American indicator and 3.8 percent for the college graduate mother indicator – but to alleviate the endogeneity concerns we also propose a sibling fixed effects strategy.[16]

neighborhoods for the purposes of school assignment and school bus routing and scheduling. While the neighborhoods vary in size, on average between 50 and 200 students live in such unit at any given time. We cannot provide specific numbers to preserve the anonymity of the school district in question.

[14] Control variables in the analysis of college outcomes include indicators for limited English proficiency, ever receiving subsidized meal in school and ever living in low-income neighborhood.

[15] We do not impose this restriction on kindergarten readiness because we do not have data for cohorts 1997 to 1999. The results are similar when we estimate the effects on redshirting, retention and test scores for all children for whom we can observe kindergarten readiness.

[16] Dickert-Conlin and Elder (2010) provide evidence that parents with different characteristics do not systematically time childbirth around school entry cutoff dates. Buckles and Hungerman (2013) provide a counter example; however, they consider seasonality across the full year rather than adjacent months comparison.

In order to implement the fixed effects we first restrict the sample to families where we observe at least two siblings in our data. Then we further require that these siblings are first two in the family and both are born in either September or August. The estimating equation becomes:

$$Y_{ij} = \delta_j + \beta Sept_{ij} + \gamma X_{ij} + \varepsilon_{ij} \tag{2}$$

where Y, $Sept$, X and ε are defined as in Equation 1 but are now additionally subscripted with j, which indexes families. In Equation 2, δ_j is a mother fixed effect that accounts for observable and unobservable characteristics that are shared by siblings and do not vary over time. Additional control in vector X is an indicator for being second born and the standard error ε is now clustered at the mother level for all outcomes. The identifying variation comes from the fact that one of the siblings is youngest and one is the oldest in their grades at school entry. Although an improvement over simple OLS the potential endogeneity concerns that this strategy cannot resolve are any form of cross-sibling reinforcing/compensatory behavior or sibling spillovers (Black et al. 2017; Landerso et al. 2017a; Qureshi 2017). We directly investigate the former one by examining redshirting and retention. The latter is beyond the scope of this analysis. However, given that we find remarkably similar academic achievement estimate across different samples and estimation strategies we think that direct sibling spillover correlated with school starting age is an unlikely source of bias in the case of Florida.

2.3 Samples

Of the singleton children in Florida who are successfully matched to education records as outlined in Section 2.1, we have 291,129 observations of children with a kindergarten readiness assessment and 794,315 observations with information on the test scores and progression through primary school. Table 1 outlines the sample sizes and the means for these samples of singletons in Panel A. We find that on average about 86 percent of the sample is deemed "kindergarten ready" with about 90 percent among September births but only 79 percent among August births. The average test score is positively selected because of the modeling decisions described in Section 2.1, but we observe that test scores are about 0.2 SD higher for September births than for August births on average. In our sample, approximately 1.4 percent of the children are redshirted but around 18 percent are retained before third grade; September-born children are dramatically less likely to have been redshirted or retained in the early grades compared to August-born children.

Panel B restricts the sample to sibling pairs where one child was born in August and the other in September. This restriction decreases the sample size significantly and because we would be restricted to families with August and September births in years when the state administered the kindergarten readiness assessment, we are unable to perform the sibling comparison analysis on kindergarten readiness. This sample selection increases the average test scores but redshirting and retention stay at similar levels. Panel C further restricts the sample to siblings that have the same observed father. This once again increases the average test scores but now also changes the redshirting and retention averages.

Table 2 presents means and sample sizes for college outcomes and juvenile incarceration from the large anonymous Florida county. September-born children are 1.5 percentage points (2.5 percent) more likely to attend college, 1.4 percentage points (4.1 percent) more likely to receive a college degree, and 1.5 percentage points (8.0 percent) more likely to graduate from a competitive institution than are August-born children. Meanwhile, they are 0.16 percentage points (21.2 percent) less likely to be incarcerated in the juvenile justice system by their 16th birthday.

3 Results

3.1 Short- and medium-run outcomes

Table 3 documents the effect of school starting age on test scores, comparing September versus August born children, for a variety of samples and specifications. The main take home point of this table is that the point estimates are quite similar regardless of the sample and/or specification. Panel A uses our entire sample of singleton births and shows the average grade 3 to 8 test score advantage of a September birth versus August birth. In Column 1 we see that the September births score 0.197 SD higher than their August counterparts. In Column 2 we add year dummies as well as a set of early health controls, including log birth weight, gestational age, indicators for prenatal care started in first trimester, congenital anomalies, abnormal conditions at birth and maternal health problems in pregnancy; and find that the point estimate barely budges – to 0.195 SD. In Column 3 we further add demographic controls which include indicators for race, ethnicity, gender, Medicaid paid birth as well as maternal education, and once again find no real difference, with an estimated effect of September birth versus August birth of 0.201 SD.

We next move to a specification in which we compare August and September births *within the same family*, by controlling for family fixed effects. We first confirm, in Panel B, that the OLS regression discontinuity estimates are essentially the same if we focus on the set of observed siblings relative to the full set of singletons; the point estimate ranges from 0.213-0.216 SD for this sample, similar to the 0.195-0.201 SD estimated for the full population of singletons. When we actually control for family fixed effects in Panel C, we find the results are extremely similar – ranging from 0.216-0.218 SD – and when we choose an even more restrictive comparison, in which we estimate sibling fixed effects regression discontinuity models when both parents are the same for both siblings (Panel D), the estimates remain essentially unchanged, ranging from 0.221-0.223 SD.[17] In summary, while one might have been concerned that unobserved family characteristics for children born in September versus August might be driving observed differences in outcomes for September versus August births, the results from Table 3 make clear that controlling for family characteristics and behavior does not substantially affect the estimated relationship between school starting age and test scores. We conclude ex post from this analysis that much of the regression discontinuity estimates in the literature are most likely not contaminated with quantitatively important family selection issues.

In Section 2.1 we have noted that our sample consists only of children who attend public schools in Florida and stay in the system at least until third grade, the first time we observe test scores. Since this sample is positively selected and the selection correlates with being born in September (Table A1) the estimates presented in Table 3 may be biased.[18] To address this problem we propose a bounding exercise where we impute either 5th or 95th percentile of tests scores to students whom we either do not match to public schools or do not observe with test scores in public schools (for example because they leave the public schools between kindergarten and commencement of testing).

[17] We do not separately estimate models in Column 3 for these specifications because the demographic variables are largely the same across births within a family. When we include time varying maternal characteristics like marital status or age at birth our results remain unchanged.

[18] We formally document this selection in Table A3, where the dependent variables are either being matched between birth and public school records or being observed with third grade test scores conditional on being merged to public school records. Since the sibling match occurred via school records, this particular analysis can only be done for the latter selection. Regardless of the specification, we find that September born children are about 2 percentage points less likely to be merged between birth and school records and are between 0.4 and 0.6 percentage points more likely to be included in the empirical sample conditional on being merged between the two data sources. The sum of these two coefficients is approximately the difference in the fraction of September births between columns 2 and 3 in Table A1.

These bounds are presented in square brackets in Table 3 and suggest that our preferred estimates are not substantively biased due to selection. The range of the bounds is also no greater than 6 percent of a standard deviation, consistently less than a third of the estimated effect in the most conservative approach.

In Figure 1, we examine the relationships found in Table 3 in more depth. In particular, we display the point estimates which come from a separate month-to-month comparison using our larger sample of singletons on test scores, as well kindergarten readiness, early retention, and redshirting. In Panel A we observe that, regardless of which month-to-month comparison we employ, the older children of the pair are more likely to be ready for kindergarten at the start of formal schooling. However, in all cases except for the September versus August comparison, the estimated differences are small, albeit often statistically significantly distinct from zero. On the other hand, in the case of the September versus August discontinuity, the difference is dramatically larger than seen elsewhere – an older-child advantage of 10 percentage points – over five times higher than the second-largest difference. For test scores, reported in Panel B, the September versus August estimate is 0.16 SD larger than second-largest difference (0.20 SD vs 0.04 SD).

Panels C and D of Figure 1 show the differential effects of being older on the probability of being redshirted (Panel C) or being retained in early grades (Panel D). Here we find that the September versus August difference in redshirting rates (5 percentage points) is more than double the next largest month to month comparison. Parents redshirt children born in both July and August but roughly twice as many August babies are redshirted than those born in July. Regarding early-grade retention (Panel D), the point estimate for the September versus August comparison is -0.151 and dwarfs any of the month by month comparisons. Therefore, Figure 1 gives us much confidence that our regression discontinuity design is accurately picking up the important age differences in our data.[19]

3.2 Heterogeneity

A majority of the previous research using US data has offered little insight in terms of heterogeneity in the school starting age estimates. Elder and Lubotsky (2009) provide some evidence of heterogeneity using data from ECLS-K, which contains a nationally representative sample of children who entered kindergarten in the 1998/99 school year. Overall, they find that higher socioeconomic status families have larger old versus young effects starting in the fall of kindergarten through the spring of eighth grade in reading and through spring of fifth grade in math. That said, they are not able to carry out this heterogeneity analysis using their preferred regression discontinuity design.[20] Cook and Kang (2016) investigate different definitions of low SES – specifically, mother being unmarried, mother being a high school dropout, and student being eligible for free or reduced-price school lunches – and find that their estimates regarding academic achievement and crime are larger in the most disadvantaged groups, but because of the nature of their study they do not investigate the effects of school starting age for relatively advantaged families. The Florida data are suited to explore heterogeneity in significantly more detail as we have access to population level data set of a highly diverse population that includes over 20 percent of African-American, Hispanic and high school dropout families. In the analysis that follows, we investigate the degree to which estimated effects of school starting age vary by race/ethnicity, maternal education, family poverty, birth weight, gestational age, school quality, and sex. The interaction between initial endowments and school

[19] Sibling fixed effects results for panels B to D are qualitatively very similar but have larger standard errors due to decreased sample sizes. This is consistent with findings reported in Table 3

[20] We are able to conduct heterogeneity analyses using the regression discontinuity design because our dataset contains over 55 times the number of students as the one used by Elder and Lubotsky (2009).

starting age, has never been studied before to our best knowledge, and appears crucial from the policy perspective given the hypothesized interaction between early childhood inputs (Cunha et al. 2010).

We present the heterogeneity results in Figures 2-6.[21] In each figure, the bar or dot represents a point estimate and includes a 95 percent confidence interval (whiskers) from our September versus August singletons regression discontinuity comparison.[22] As seen in Figure 2, Panel A, the September-August difference in kindergarten readiness is much lower for high-SES families than for low-SES families (whether measured by family income proxied by Medicaid payment or maternal education groups); and much lower for white families than for minority families.[23] These are exactly the groups that also experience higher redshirting rates. On the other hand, differences in readiness are comparatively low for higher-birthweight infants relative to lower-birthweight infants or full-term infants relative to premature or post-term infants suggesting no interaction between initial health endowments and age at the start of education (see Figures 3 and 4).

Remarkably, as seen in Figures 2-5, the estimated effects of school starting age on test scores are highly similar across a wide range of SES groups as well as a wide range of initial infant health, or a wide range of school quality (as measured by the gain scores of the school calculated by the Florida Department of Education for the purposes of school accountability).[24] These findings indicate that school starting age affects children's test scores by essentially the same amount – despite the fact that different groups of families have babies with different average health at birth or academic achievement and are differentially proactive regarding how they attempt to remediate their young-for-grade children.

Differences in early family remediation behaviors can help to explain why we document considerable heterogeneity in kindergarten readiness but not in third grade test scores by different family background groups. We postulate that remediation behaviors might be partially responsible for the presence of heterogeneity at the start of school but not in subsequent test scores as we observe that high-SES families are more likely to redshirt their August-born children, while children from low-SES families are more likely to be retained in early grades. Importantly while redshirting has the potential of affecting both kindergarten readiness and subsequent test scores by the nature of school retention it happens only after a child starts schooling, and thus cannot have an effect on kindergarten screening results. This difference in timing is consistent with the pattern observed in the data, and the two approaches to remediating relatively young children may be the cause of the sharply reducing SES-age profile for August-born versus September-born children by third grade. Later in this paper we provide some suggestive evidence regarding the potential efficacy of these remediation strategies.

Exploring heterogeneity further, we look to the students' sex. Boys are redshirted more often than girls (Bassok and Reardon 2013; Schanzenbach and Howard 2017), implying that many families think that school starting age is more relevant for their sons than for their daughters. In Figure 6, we graph the point estimate for males and females in our sample. In terms of kindergarten readiness, we find that September males have a marginally larger age disadvantage than females. However, in terms of averaged test scores, we are unable to statistically distinguish between male and female difference in our September versus August estimates – they are equally as big, around 0.2

[21] The point estimates and associated standard errors can be found in Appendix Tables A4 to A7.

[22] Our sibling fixed effects heterogeneity results are again qualitatively similar; however, due to small sample sizes we often lose statistical power.

[23] Elder and Lubotsky (2009) also find significant heterogeneity during the fall of kindergarten but they find larger age effects for the children from higher socioeconomic status families, which is at odds with our estimates.

[24] We are unable to explore differences in kindergarten readiness or redshirting practices stratified by school quality as these two outcomes are measured at the very beginning of schooling, and thus cannot be affected by the quality of school that a child attends in the first grade.

SD. At the same time, there are significant gender differences in behaviors of parents and schools in terms of redshirting and retention. Male August babies are significantly more likely to be redshirted than female August babies, perhaps due to "conventional wisdom" regarding gender differences in maturity, or perhaps due to the fact that August-born boys are somewhat less ready to start school than are August-born girls. August-born boys are also differentially more likely to be retained in early grades (relative to their September-born counterparts) than are August-born girls. It is unclear if the redshirting or retention caused the male test scores to increase, decrease, or have no effect at all. What we do know is that despite higher levels of both retention and redshirting, males and females experience similar test score gaps.

We further examine the heterogeneity by socioeconomic status and gender by providing each heterogeneity estimate separately for boys and girls (see Autor et al. (2016a) and Autor et al. (2016b) for an in depth exploration of gender-SES gaps in Florida). The results can be found in Table 4. In Panel A, we find that across most categories, the kindergarten readiness gap between September versus August is larger for males than females. However, this difference is only statistically significant in three out of seven groups. When examining the average test score gap in Panel B, we find that the test score gap is similar between males and females except for the children with college educated mothers and mothers who were not on Medicaid. In these cases, the test score gap is actually larger for females. We find that the August born males are redshirted and retained more in all categories but the magnitude of redshirting is substantially higher for the boys with mothers who are college graduates, non-Medicaid, or white. These facts together indicate that the increased prevalence of redshirting might help to boost test scores as we have seen that these males are also the children that have a smaller September-August test score gap.

Summarizing our heterogeneity analysis, we observe that there exists very little heterogeneity in age effects on test scores across a substantial array of different child, family, and school dimensions, despite pronounced age effects in kindergarten readiness. These differences could be due to the consequences of various remediation efforts. In Section 3.4 of this paper we attempt to uncover whether remediation efforts could be responsible for some of these patterns of results.

3.3 Longer-run outcomes – juvenile detention and college outcomes

In a large anonymous county in Florida, we are able to explore the school starting age effects for longer-run outcomes such as attending college, obtaining a college degree, and obtaining a college degree from a competitive/selective institution as defined by Barron's. We also study the effects of school starting age on the probability of juvenile incarceration before a child's 16th birthday. In addition, we explore the heterogeneity of these results by childhood socioeconomic status proxied by micro-neighborhood affluence and by student race/ethnicity.[25] While we lack the statistical power to compare siblings in this context, we can still estimate regression discontinuity models, as before.

Table 5 displays the βs from Equation 1. We observe that on average September-born children are 1.3 percentage points (2.1 percent) more likely to attend college, 1.1 percentage points (3.3 percent) more likely to graduate from college, and 1.3 percentage points (7.2 percent) more likely to graduate from a competitive/selective college than are August-born children, all else equal.[26] While not statistically distinct from zero at conventional levels, we also observe that September-born children are 0.15 percentage points (15.4 percent) less likely to be incarcerated for juvenile

[25] Specifically, we stratify micro-neighborhoods by the fraction of a child's immediate neighbors who are eligible for free or reduced-priced lunches.

[26] Dobkin and Ferreira (2010) using a RD analysis do not find statistically significant effects of school entry laws on college completion on average in samples from Texas or California. However, they do find a marginally significant effect for white students in Texas but not in California.

crime before their 16th birthday. These results indicate that being relatively young at school entry has some lasting implications.

Our heterogeneity analyses indicate that white students tend to experience larger September-versus-August effects on these longer-run outcomes than do minority students: In the case of college attendance and college completion, the estimated benefits of September birthdays are concentrated entirely among white families. For selective college completion, the estimated effects for white and minority families appear to be about the same, with effect sizes larger for children from minority families, though these are imprecisely estimated. The estimated reductions in juvenile incarceration associated with September births are also concentrated among white rather than minority families. Regarding SES stratifications, the largest estimated benefits of September birthdays on college attendance and completion are apparently enjoyed by medium-SES families, although both medium-SES and high-SES September-born children experience increased likelihoods of graduation from competitive/selective colleges. We lack the statistical power to discern SES differences in juvenile incarceration.

In contrast to some studies (e.g. Fertig and Kluve (2005) and Black et al. (2011)) and in line with other papers (e.g. Bedard and Dhuey (2006) and Fredriksson and Ockert (2014)), these estimates provide evidence that a child's age at school entry can have long-lasting effects on children's life chances. It is striking that many of our estimated effects are concentrated in the white and middle SES families (or even high-SES, in the case of selective college completion), a finding similar to that of Dobkin and Ferreira (2010), who find marginally significant effects for white individuals in Texas. These are the same people who frequently have a lot of ability to mitigate issues related to disadvantage in the academic setting – but age at school entry does not seem to be one of these issues that they can mitigate easily. However, it is interesting that the age at school entry effect persist for these particular groups of individuals despite similar levels across different groups in test scores in the earlier years. It may be due to the fact that a higher percentage of high SES children attend college than lower SES children so the apparent margin for impact of age at school entry is larger among relatively more affluent families.

3.4 Exploratory analysis: Potential consequences of redshirting and early grade retention

Providing causal evidence on the effect of redshirting is difficult as children who are being redshirted undoubtably come from families that are different on both observables and unobservables. For example, in our sample, families with college-educated mothers are more likely to redshirt their children in comparison to families where the mothers did not complete college education. Thus, it is difficult to disentangle the act of redshirting from the observable and unobservable qualities of these families. We provide some evidence in Section 3.2 that redshirting might potentially increase the test scores of those being redshirted - primarily males from higher socioeconomic status families. We next provide additional associational evidence on this relationship.

Schools and school districts are often given leeway on the rules and regulations set in their district regarding who is allowed to be redshirted. Some districts allow for large levels of redshirting whereas others do not. Also, different "parenting cultures" and even daycare prices may affect redshirting practices which in turn may affect redshirting levels across school districts and within school districts over time. As a consequence, across the 65 county-level school districts in Florida redshirting rates vary from 0 to 8.5 percent, and August-birth redshirting rates range from 0 to 50 percent.[27] School

[27] We exclude two counties from this analysis because we do not observe children's place of residence at birth for those born in 1994 and 1995 in these counties. These two counties constitute 1.5% of the full population of births in years 1996 to 2000. Our results are fundamentally unchanged when we use all 67 school districts and limit birth

districts vary dramatically in terms of early-grade retention rates as well. Early-grade retention rates range from 9.1 to 48.3 percent across the 65 school districts, and early-grade retention rates for August births range from 0 to 100 percent.[28] This variation is not due to observed background differences: If we were to predict redshirting and early-grade retention using the variables observed on a child's birth certificate, we would have only expected to see a range of 1.0 to 2.6 in county-level differences in redshirting rates and a range of 14.4 to 26.7 in county-level differences in early-grade retention rates.[29] Finally, districts that pursue one policy or practice do not necessarily pursue the other policy or practice. For example, district-cohort correlation between redshirting and retention rates is 0.657.

In Table 6, we examine the relationships between school district-level differences in the rates of redshirting as well as early-grade retention and test scores. We collapse the data at year of birth×school district×September births level. In the first column of Panel A, we regress average test scores on the fraction of children redshirted, an indicator for a September birth and the interaction of these two variables. We also include school district and cohort fixed effects and cluster our standard errors at the school district level. We find that the percent redshirted is positively related to the average test score level but that the interaction between percent redshirted and being born in September is negatively related. This indicates that the school districts that have higher amounts of children who are redshirted have lower September/August test score gaps. In the first column of Panel B, we regress average test scores on the fraction of children retained by the school in early grades. Here we find that retention is negatively related to test scores but that the interaction is also negatively related. This implies that school districts that have higher levels of retention in early grades also have lower old versus young test score gaps. This evidence is necessarily only suggestive because despite the school district and cohort fixed effects there may still be unobservable variables that affect both the level of redshirting/retention and the level of test scores. Nonetheless, this district level evidence paired with previous individual level analyses gives us confidence to lean in the direction of saying that school districts where redshirting and early grade retention are more prevalent also have smaller September-August gaps in test scores.

We further investigate this question by considering heterogeneity in *whose* test scores are differentially related to school district-level redshirting and early grade retention rates.[30] The remaining columns of Table 6 help to tell this story. While only correlational in nature, we find that in school districts where one remediation strategy is especially prevalent, September-August performance gaps tend to be lower for the demographic and socio-economic groups that in general are less likely to experience that type of remediation. In the case of redshirting, the largest reductions in the September-August performance gaps associated with district-level redshirting rates are for children with high school dropout mothers and for children who were born in poverty, as indicated by Medicaid-funded births. In the case of early grade retention, the largest reductions in the September-

cohorts to 1996 to 2000, when we observe location of birth for the entire state.

[28]Some of the school districts in our sample are very small and have less than 10 students born in a given year and month. If we restrict our sample to counties with at least 50 August births in each year we are left with 29 school districts, and then the August redshirting rates range from 1.3 to 18.1 while August early retention rates range from 12.1 to 44.3.

[29]We regress at individual level indicator for being redshirted or early retained on infant gender, month of birth dummies, birth weight, gestational age, dummies for congenital anomalies and abnormal conditions at birth as well as on mom's race, ethnicity, education, foreign born status, medicaid paid birth, health problems and start of prenatal care in first trimester. Then, we use coefficients from this regression to predict values of the dependent variables and collapse them at school district (65 districts) and year level (7 years).

[30]Here, we limit our analysis to school districts with at least five children in each cell (year of birth by September birth heterogeneity groups). This restriction yields unbalance repeated-cross section of observations. Retention results are similar when we impose full panel restriction across heterogeneity dimensions and years while the results for redshirting become less precisely estimated.

August performance gaps associated with district-level early grade retention rates are for children with high school graduate and college graduate mothers; for those whose births were *not* funded by Medicaid; and for white students rather than minority students. This pattern of findings provides further evidence that while redshirting and early grade retention are remediation tools that could have negative consequences, such as those described by Schanzenbach and Howard (2017), there are potential vehicles for remediation inside and outside of the school system.

4 Conclusions

In this paper, we document, using matched administrative data from the state of Florida, the most robust to date evidence on the effects of school starting age on children's cognitive development. The regression discontinuity approach as well as the month-to-month within family sibling fixed effects comparison where we control for all the time invariant genetic, birth endowments and family characteristics show that September born children benefit developmentally in comparison to August born children. Our findings are very similar irrespective of the empirical approach chosen which suggests that much of the regression discontinuity estimates in the literature thus far are most likely not contaminated with quantitatively important family selection issues.

We do find some heterogeneity in terms of kindergarten readiness but we also document a striking lack of heterogeneity in these effects by student, maternal, and school characteristics for test scores. At the same time, we observe different compensatory behaviors targeted towards children from different socioeconomic statuses who are youngest in their schooling cohort. While the more affluent families tend to redshirt their children to give them competitive advantage, families that are unable to do this - either due to lack of awareness or resources - are surrogated by the schooling system, which retains their children in grades prior to testing. This differential remediation also helps explaining why we find larger kindergarten readiness gaps for lower SES children that then vanish at the time of testing. Namely, since low SES children are not redshirted but rather retained there is no scope for retention to affect children's cognitive development prior to the start of schooling. Together, both of these mechanism seem to be effective because children coming from different socio-economic backgrounds end up at roughly the same educational levels at the time of testing irrespective of the affluence.

In one anonymous county in Florida we are also able to document significant effects on longer-run outcomes that show, despite some literature finding that test-score effects fade out of age by later grades (Elder and Lubotsky 2009), that the age at school entry may still impact children through their lifetime via their educational opportunities. Contrary to the test score results, we document significant heterogeneity that can be found in these longer-run outcomes by race/ethnicity and socioeconomic status. We find that the estimated effects are concentrated in the white, middle of the distribution of socioeconomic status children.

Finally we explore whether the relationship between remediation techniques and test scores estimated at individual level translated into policy relevant district-level variation. We show that the percent of children redshirted is positively related to the average test score level but that the interaction between the percent redshirted and being the oldest in a cohort is negatively related. At the same time, retention is negatively related to test scores but the interaction between retention and being the oldest in a cohort is also negatively related. Together, these findings indicate that school districts where redshirting and early grade retention are higher have smaller relative age gaps in test scores.

References

Attar, Itay and Danny Cohen-Zada, "The effect of school entrance age on educational outcomes: Evidence using multiple cutoff dates and exact date of birth," IZA Discussion Paper 10568, 2017.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman, "Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes," NBER Working Paper 22267, 2016.

_ , _ , _ , _ , and _ , "School Quality and the Gender Gap in Educational Attainment," *American Economic Review Papers and Proceedings*, 2016, *106* (5), 289–295.

Barker, David, "The fetal and infant origins of adult disease," *BMJ*, 1990, *301* (6761), 1111.

Bassok, Daphna and Sean Reardon, "Academic redshirting in kindergarten: Prevalence, patterns, and implications," *Educational Evaluation and Policy Analysis*, 2013, *35* (3), 283–297.

Bedard, Kelly and Elizabeth Dhuey, "The persistence of early childhood maturity: International evidence of long-run age effects," *Quarterly Journal of Economics*, 2006, *121* (4), 1437–1472.

Bharadwaj, Prashant, Petter Lundborg, and Dan-Olof Rooth, "Birth weight in the long run," *Journal of Human Resources*, 2017, *forthcoming.*

Black, Sandra, Paul Devereux, and Kjell Salvanes, "From the cradle to the labor market? The effect of birth weight on adult outcomes," *Quarterly Journal of Economics*, 2007, *122* (1), 409–439.

_ , _ , and _ , "Too young to leave the nest? The effects of school starting age," *Review of Economics and Statistics*, 2011, *93* (2), 455–467.

_ , Sanni Breining, David Figlio, Jonathan Guryan, Helena Nielsen Skyt, Jeffrey Roth, and Marianne Simonsen, "Sibling spillovers," NBER Working Paper 23062, 2017.

Buckles, Kasey and Daniel Hungerman, "Season of birth and later outcomes: Old questions, new answers," *Review of Economics and Statistics*, 2013, *95* (3), 711–724.

Cascio, Elizabeth and Diane Whitmore Schanzenbach, "First in the class? Age and the education production function," *Education Finance and Policy*, 2016, *11* (3), 225–250.

Cook, Philip and Songman Kang, "Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation," *American Economic Journal: Applied Economics*, 2016, *8* (1), 33–57.

Crawford, Claire, Lorraine Dearden, and Costas Meghir, "When you are born matters: The impac of date of birth on child cognitive outcomes in England," London: Centre for the Economics of Education, 2007.

Cunha, Flavio, James Heckman, and Susanne Schennach, "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 2010, *78* (3), 883–931.

Currie, Janet and Hannes Schwandt, "Within-mother analysis of seasonal patterns in health at birth," *Proceedings of the National Academy of Sciences*, 2013, *110* (30), 12265–12270.

Datar, David, "Does delaying kindergarten entrance give children a head start?," *Economics of Education Review*, 2006, *25* (1), 43–62.

Deming, David and Susan Dynarski, "The Lengthening of Childhood," *Journal of Economic Perspectives*, 2008, *22* (3), 71–92.

Dhuey, Elizabeth and Stephen Lipscomb, "What makes a leader? Relative age and high school leadership," *Economics of Education Review*, 2008, *27* (2), 173–183.

_ and _ , "Disabled or young? Relative age and special education diagnoses in schools," *Economics* *of Education Review*, 2010, *29* (5), 857–872.

Dickert-Conlin, Stacy and Todd Elder, "Suburban legend: School cutoff dates and the timing of births," *Economics of Education Review*, 2010, *29* (5), 826–841.

Dobkin, Carlos and Fernando Ferreira, "Do school entry laws affect educational attainment and labor market outcomes?," *Economics of Education Review*, 2010, *29* (1), 40–54.

Du, Qianqian, Huasheng Gao, and Maurice Levi, "The relative-age effect and career success: Evidence from corporate CEOs," *Economics Letters*, 2012, *117* (3), 660–662.

Elder, Todd and Darren Lubotsky, "Kindergarten entrance age and children's achievement: Impacts of state policies, family background, and peers," *Journal of Human Resources*, 2009, *44* (3), 641–683.

Evans, William, Melinda Morrill, and Stephen Parente, "Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children," *Journal of Health Economics*, 2010, *29* (5), 657–673.

Fertig, Michael and Jochen Kluve, "The effect of age at school entry on educational attainment in Germany," IZA Discussion Paper 1507, 2005.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth, "The effects of poor neonatal health on children's cognitive development," NBER Working Paper 18846, 2013.

_ , _ , _ , and _ , "The effects of poor neonatal health on children's cognitive development," *American Economic Review*, 2014, *104* (12), 3921–3955.

_ , _ , _ , and _ , "Long-term cognitive and health outcomes of school-aged children who were born late-term vs. full-term," *JAMA Pediatrics*, 2016, *170* (8), 758–764.

Fredriksson, Peter and Bjorn Ockert, "Life-cycle effects of age at school start," *Economic Journal*, 2014, *124* (579), 977–1004.

Garfield, Craig, Krzysztof Karbownik, Karna Murthy, Gustave Falciglia, Jonathan Guryan, David Figlio, and Jeffrey Roth, "Educational Performance of Children Born Prematurely," *JAMA Pediatrics*, 2017, *171* (8), 1–7.

Kawaguchi, Daiji, "Actual age at school entry, educational outcomes, and earnings," *Journal of the Japanese and International Economies*, 2011, *25* (2), 64–80.

Landerso, Rasmus, Helena Skyt Nielsen, and Marianne Simonsen, "How going to school affects the family," Department of Economics Aarhus University, 2017.

_ , _ , and _ , "School starting age and the crime-age profile," *Economic Journal*, 2017, *forthcoming.*

Lubotsky, Darren and Robert Kaestner, "Do skills beget skills? Evidence on the effect of kindergarten entrance age on the evolution of cognitive and non-cognitive skill gaps in childhood," *Economics of Education Review*, 2016, *53*, 194–206.

McAdams, John, "The effect of school starting age policy on crime: Evidence from U.S. microdata," *Economics of Education Review*, 2016, *54*, 227–241.

McEwan, Patrick and Joseph Shapiro, "The benefits of delayed primary school enrollment. Discontinuity estimates using exact birth dates," *Journal of Human Resources*, 2008, *43* (1), 1–29.

Morrow, Richard, Jane Garland, James Wright, Malcolm Maclure, Suzanne Taylor, and Colin Dormuth, "Influence of relative age on diagnosis and treatment of attention-deficit/hyperactivity disorder in children," *CMAJ*, 2012, *184* (7), 755–762.

Ozek, Umut, "Hold back to move forward? Early grade retention and student misbehavior," *Education Finance and Policy*, 2015, *10* (3), 350–377.

Puhani, Patrick and Andrea Weber, "Does the early birth catch the worm? Instrumental variables estimates of early educational effects of age of school entry in Germany," *Empirical Economics*, 2007, *32* (2), 359–386.

Qureshi, Javaeria, "Siblings, teachers and spillovers in academic achievement," *Journal of Human Resources*, 2017, *forthcoming.*

Schanzenbach, Diane Whitmore and Stephanie Larson Howard, "Season of birth and later outcomes: Old questions, new answers," *Education Next*, 2017, *17* (3), 18–24.

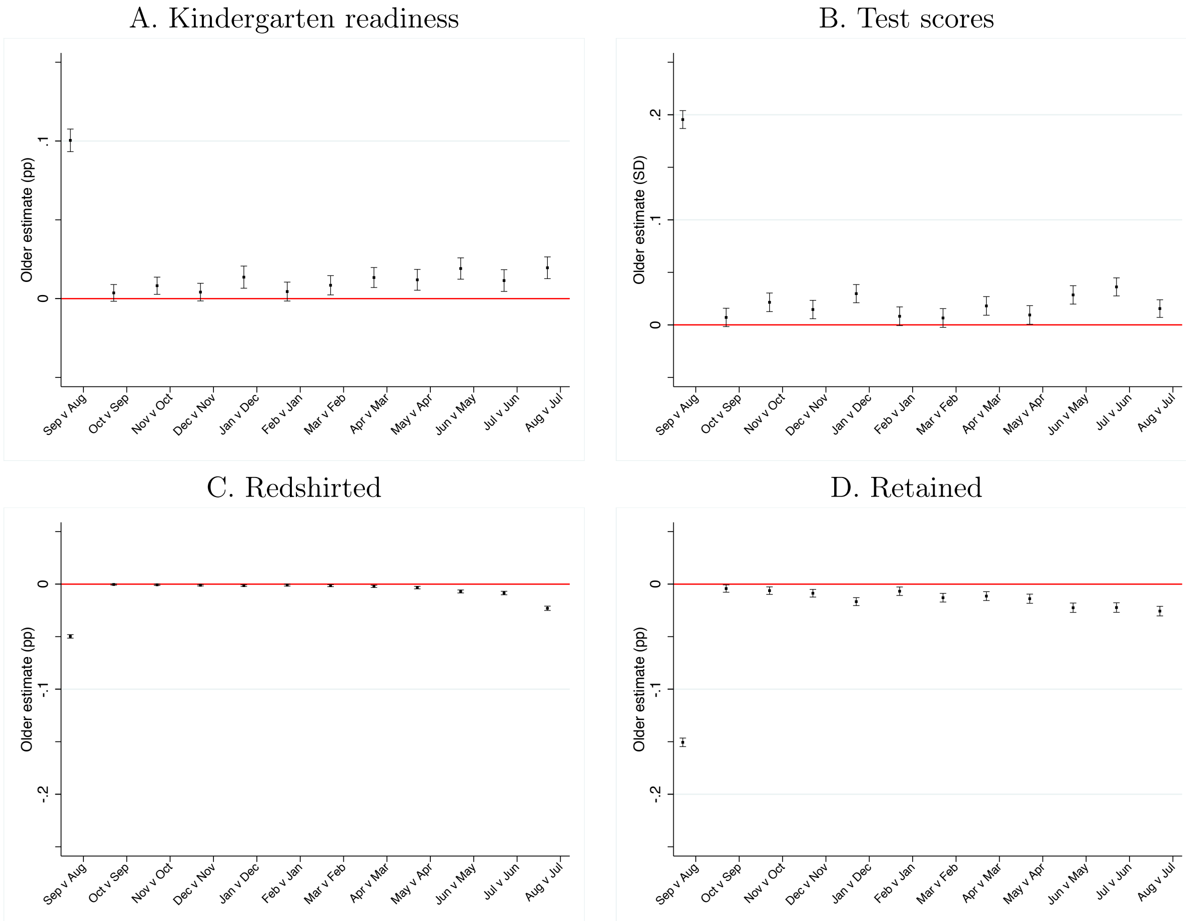
Schwerdt, Guido, Martin West, and Marcus Winters, "The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida," NBER Working Paper 21509, 2015.

Smith, Justin, "Can regression discontinuity help answer an age-old question in education? The effect of age on elementary and secondary school achievement," *The B.E. Journal of Economic Analysis and Policy*, 2009, *9* (1), 1–30.

Weil, Elizabeth, "When should a kid start kindergarten," New York Times June 3 2007.

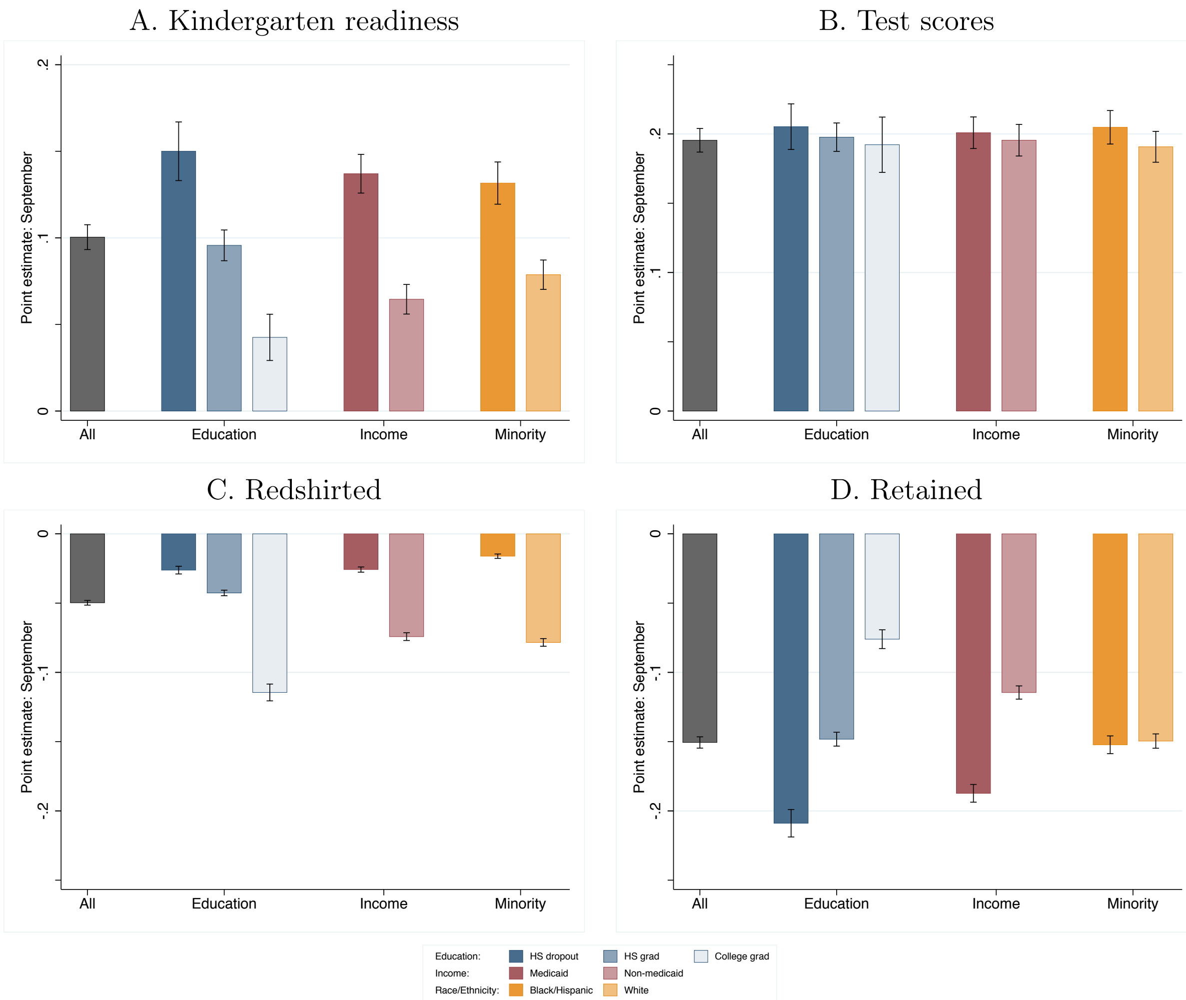
Figures and Tables

Figure 1: Estimates of school starting age (month-by-month)



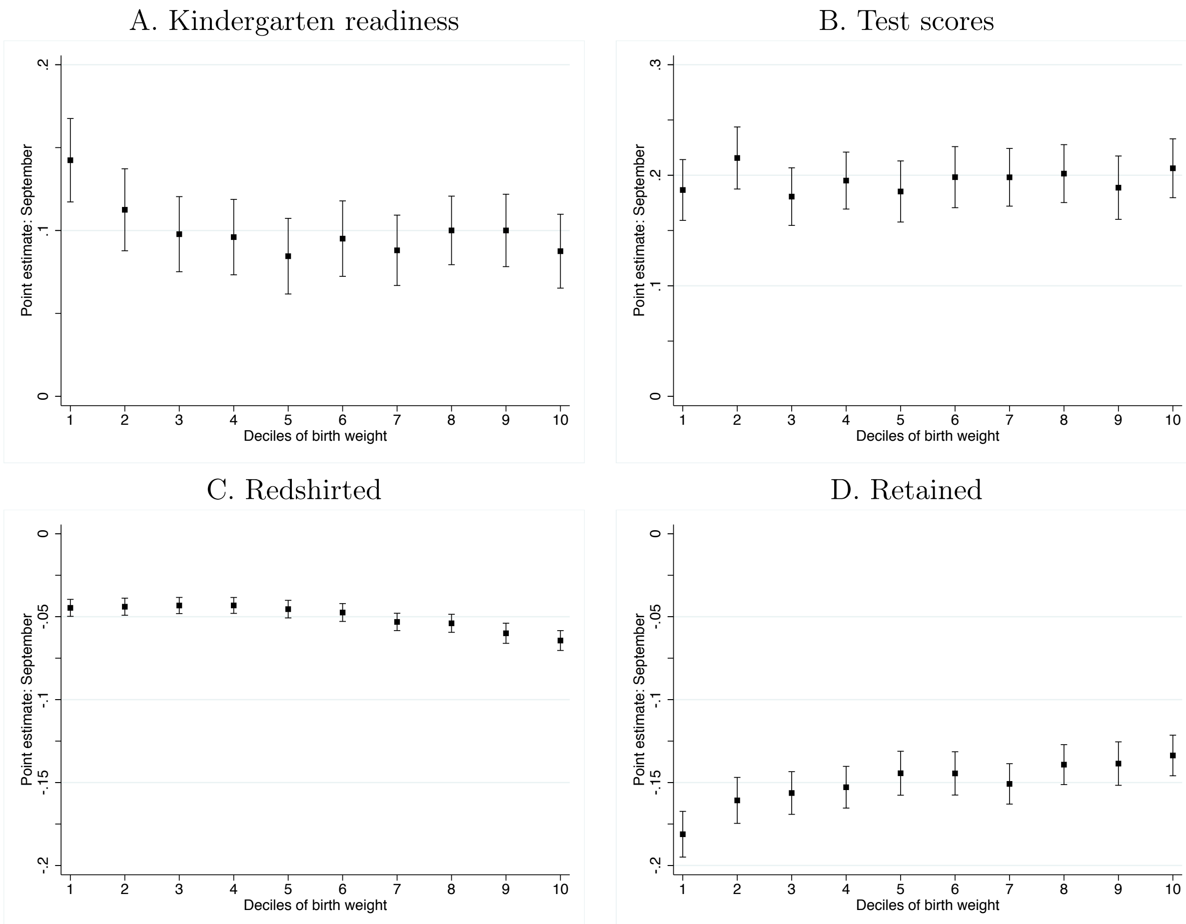
Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate presents month-to-month comparison with 95% confidence intervals. Panel A presents results for kindergarten readiness, Panel B for pooled math and reading test scores in grades 3 to 8, Panel C for probability of being redshirted and Panel D for school retention. Kindergarten readiness excludes cohorts 1997 to 1999 due to missing data. Redshirting is defined as indicator variable that equals to one if a child has a higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Control variables include: cohort indicators, log birth weight, gestational age, indicator for starting prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health problems. Heteroskedasticity robust standard errors in panels A, C and D and clustered at individual level in Panel B.

Figure 2: Heterogeneity by socioeconomic status (August vs. September)



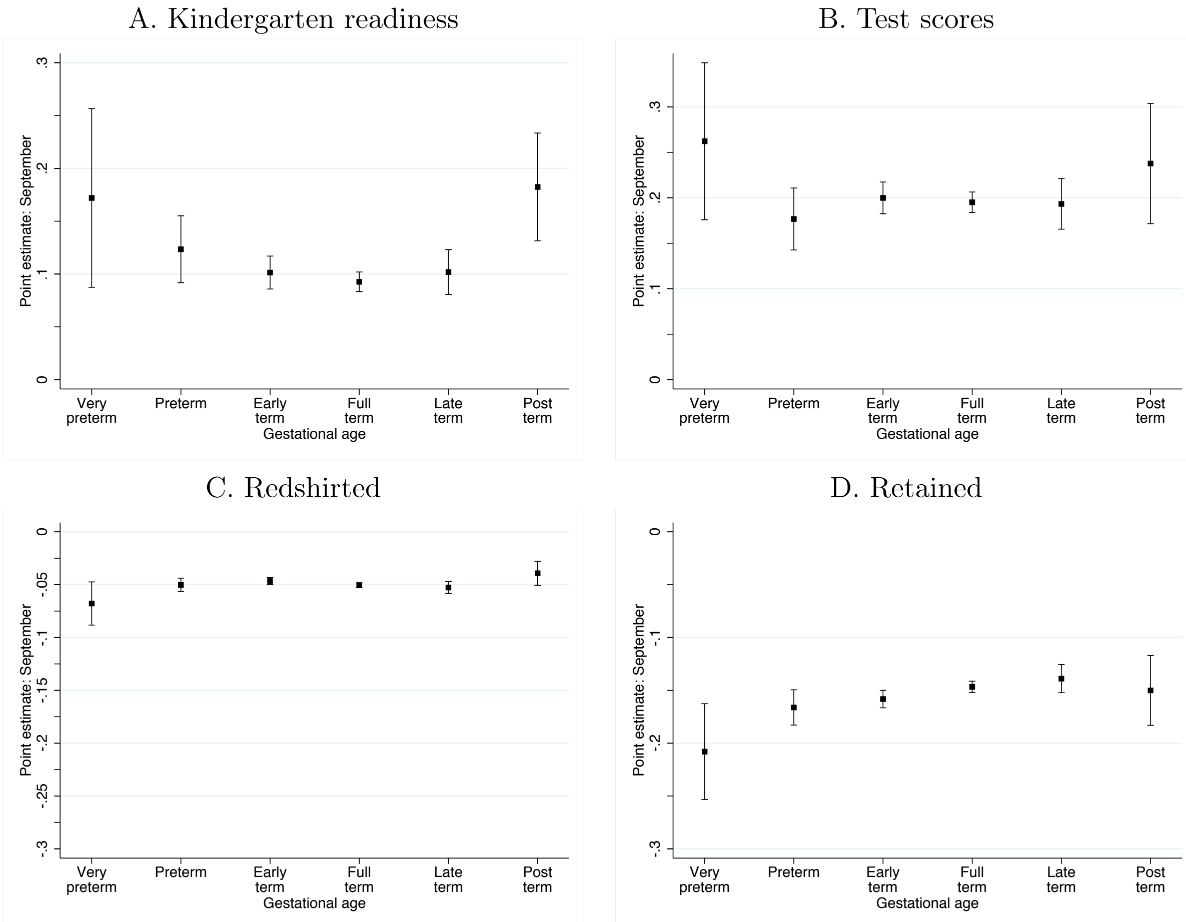
Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison with 95% confidence interval. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). Black bars present average estimates akin to those in Figure 1; blue bars present heterogeneity by maternal education, maroon bars present heterogeneity by medicaid status which is proxy for income; and orange bars present heterogeneity by race and ethnicity where minority is defined as either African-American or Hispanic. For definitions and control variables see Figure 1. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Figure 3: Heterogeneity by birth weight (August vs. September)



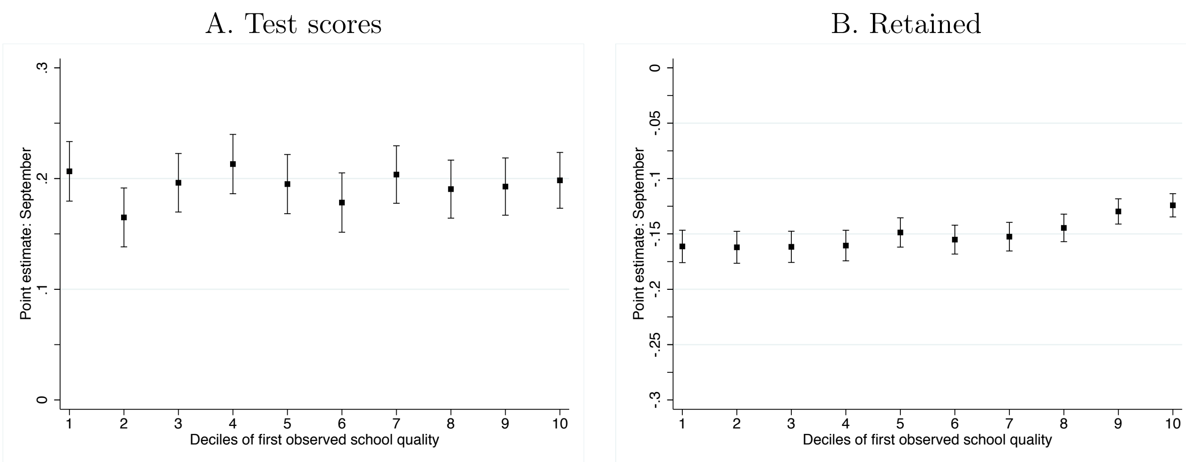
Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison for each decile of birth weight with 95% confidence interval. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). For definitions and control variables see Figure 1. These regressions exclude log birth weight and gestational age. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Figure 4: Heterogeneity by gestational age (August vs. September)



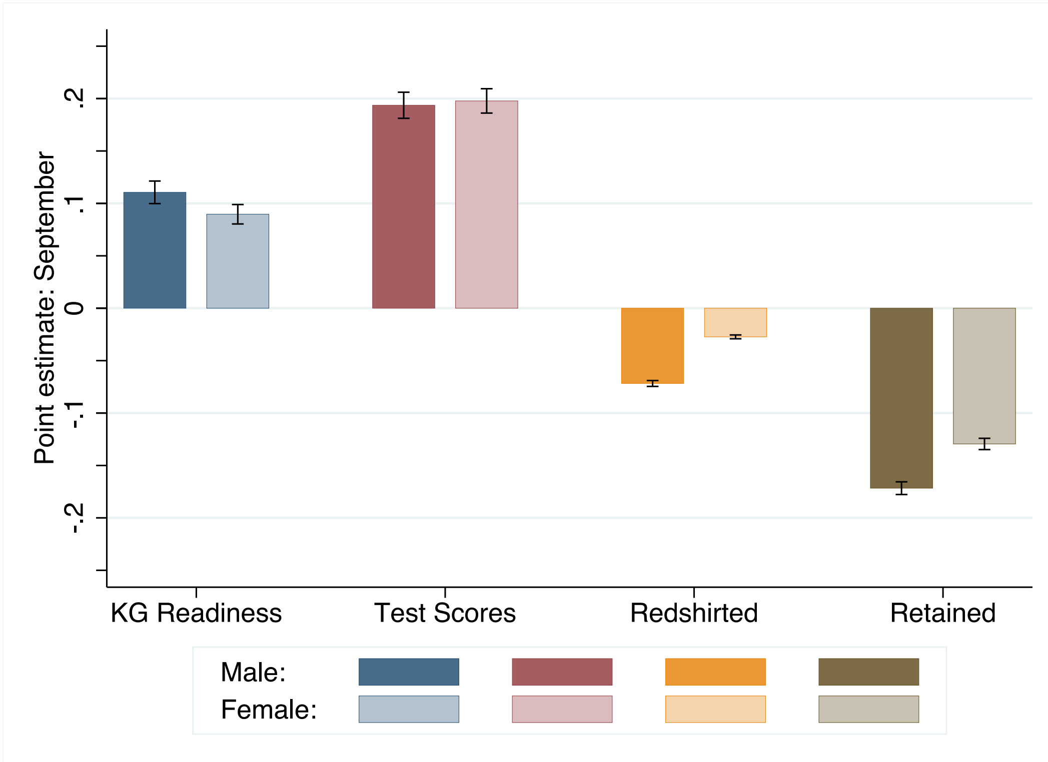
Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison for each gestational age group with 95% confidence interval. Gestational age groups are defined as follows: very preterm - below 32 weeks, preterm - 32 to 36 weeks, early term - 37 to 38 weeks, full term - 39 to 40 weeks, late term - 41 weeks, and post term - above 41 weeks. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). For definitions and control variables see Figure 1. These regressions exclude log birth weight and gestational age. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Figure 5: Heterogeneity by school quality (August vs. September)



Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison for each decile of contemporaneous school quality with 95% confidence interval. Outcomes are: pooled math and reading test scores in grades 3 to 8 (Panel A) and school retention (Panel B). For definition of school quality see Section 2.1 while for definitions and control variables see Figure 1. Heteroskedasticity robust standard errors for being redshirted and retained while clustered at individual level for test scores.

Figure 6: Heterogeneity by gender (August vs. September)



Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison with 95% confidence interval. Navy bars present results for kindergarten readiness, maroon bars for pooled math and reading test scores in grades 3 to 8, orange bars for being redshirted and olive bars for school retention. Darker bars present estimates for males and lighter for female students. For definitions and control variables see Figure 1. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Table 1: Descriptive statistics: early schooling outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Kindergarten readiness			Test scores in grades 3 to 8			Redshirted			Retained before third grade		
	Births											
	All	September	August	All	September	August	All	September	August	All	September	August
Panel A: Sample of singletons												
Mean	0.857	0.896	0.791	0.074	0.165	-0.032	0.014	0.003	0.052	0.183	0.124	0.275
N	291,129	25,139	26,305	794,315	67,997	71,214	794,315	67,997	71,214	794,315	67,997	71,214
Panel B: Sample of siblings												
Mean		N/A		0.185	0.254	0.038	0.012	0.003	0.071	0.16	0.113	0.242
N				136,814	1,092	1,092	136,814	1,092	1,092	136,814	1,092	1,092
Panel C: Sample of siblings with the same father												
Mean		N/A		0.371	0.456	0.233	0.013	0.000	0.097	0.113	0.083	0.184
N				95,266	735	735	95,266	735	735	95,266	735	735

Note: Sample is based on all singleton births between 1994 and 2000. Table 1 present means and sample sizes for the four outcome variables used in the analysis: kindergarten readiness (column 1 to 3); pooled math and reading test scores in grades 3 to 8 (columns 4 to 6), probability of being redshirted (columns7 to 9) and probability of being retained in school (columns 10 to 12). Kindergarten readiness excludes cohorts 1997 to 1999 due to missing data. Redshirting is defined as indicator variable that equals to one if a child has a higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Columns (1), (4), (7) and (10) present values for the full population; columns (2), (5), (8) and (11) for students born in September while columns (3), (6), (9) and (12) for students born in August. Panel A presents values in population of singletons, panel B for the sample of siblings and panel C for sample of siblings where the father is identical and known. In sibling sample for September and August columns we require that among the two siblings one is born in August and the other in September or vice versa.

Table 2: Descriptive statistics: longer-run schooling outcomes

	(1) All	(2) September	(3) August	(4) All	(5) September	(6) August
	A. Ever attended college			B. Obtained college degree		
Mean	0.609	0.611	0.596	0.330	0.338	0.324
Observations	95,627	8,113	8,810	95,627	8,113	8,810
	C. Obtained college degree from competitive/selective institution			D. Incarcerated in juvenile justice system by age 16 (indicator x 100)		
Mean	0.180	0.188	0.173	0.969	0.740	0.897
Observations	95,627	8,113	8,810	95,627	8,113	8,810

Note: Sample is based on all singleton births from one unnamed county in Florida for years 1981 to 1990. Outcome variables are: ever attending college (Panel A); graduating with a degree (Panel B); graduating with a degree from a competitive/selective institution defined as those rated "competitive" or above as recorded in the NCES-Barron's Admissions Competitive Index Data Files for 2004 (Panel C); and being in juvenile detention center by age 16 (Panel D). Columns (1) and (4) present mean values and sample sizes for the full population; columns (2) and (5) for students born in September; while columns (3) and (6) for students born in August.

Table 3: Effects of school starting age (August vs. September) - comparison of different econometric models

	(1)	(2)	(3)
VARIABLES		Grade 3 to 8 pooled test scores	
		Panel A: Singletons (OLS)	
September birth	0.197***	0.195***	0.201***
	(0.004)	(0.004)	(0.004)
	[0.180 to 0.234]	[0.180 to 0.235]	[0.180 to 0.238]
Observations		730,675	
Number of children		139,211	
		Panel B: Siblings (OLS)	
September birth	0.216***	0.213***	0.216***
	(0.025)	(0.026)	(0.025)
	[0.212 to 0.223]	[0.216 to 0.225]	[0.216 to 0.225]
Observations		10,910	
Number of sibling pairs		1,092	
		Panel C: Siblings (FE)	
September birth	0.216***	0.218***	
	(0.025)	(0.025)	
	[0.212 to 0.223]	[0.216 to 0.218]	N/A
Observations	10,910		
Number of sibling pairs	1,092		
		Panel D: Siblings with same parents (FE)	
September birth	0.223***	0.221***	
	(0.029)	(0.029)	
	[0.216 to 0.234]	[0.219 to 0.227]	N/A
Observations	7,476		
Number of sibling pairs	735		
Year dummies		X	X
Early health controls		X	X
Demographic controls			X

Note: Sample is based on all singleton births between 1994 and 2000. All estimates come from August vs. September comparison. Samples are: universe of singletons (Panel A); siblings born one in each month (Panels B and C) and siblings born one in each month where the father is know and the same across the two births (Panel D). OLS regressions in Panels A and B while sibling fixed effects regressions in Panels C and D. Column (1) does not include any controls; column (2) controls for cohort dummies, log birth weight, gestational age, indicator for start of prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health at birth (it also includes infant gender in sibling FE analyses - panels C and D); column (3) further controls for maternal education, Medicaid status at birth, race, ethnicity and gender. Standard errors clustered at individual level. Square brackets in this Table present estimates from a bounding exercise that we perform to address selection into the estimation sample discussed in Section 2.1. In each case, we impute either the 5th or 95th percentile of test scores for children whom we observe without test scores. The imputed percentiles are computed separately for each year of birth, month of birth and grade in school so that we can account for the fact that later born children do not reach middle school grades by the end of our test scores data span. In particular, we do not impute test scores in grade 8 for children born in 2000 and in September of 1999; and we do not impute grade 7 for children born in September 2000. In panel A we impute scores for all children born in Florida who do not make it to our empirical sample while in panels B to D we do it conditional on being observed in public school because only for this subsample we can identify siblings. The sample sizes for these bounding exercises are 1,231,791 in panel A; 16,350 in panels B and C; and 11,362 in panel D. If we apply the restriction of matching into public schools (panels B to D) to selection analysis in panel A then the estimated bounds are [0.194 to 0.210], [0.193 to 0.210] and [0.196 to 0.215] for columns (1) to (3), respectively, while the sample size is 996,057.

Table 4: Effects of school starting age (August vs. September) - differential effects for boys by maternal socioeconomic characteristis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Maternal education			Income		Race/Ethnicity	
VARIABLES	HS dropout	HS grad	College grad	Medicaid	Non-medicaid	Black/Hispanic	White
				Panel A: Kindergarten readiness			
September effect for boys	0.145***	0.111***	0.052***	0.145***	0.078***	0.137***	0.094***
	(0.013)	(0.007)	(0.010)	(0.009)	(0.007)	(0.009)	(0.006)
September effect for girls	0.152***	0.079***	0.034***	0.128***	0.050***	0.125***	0.063***
	(0.012)	(0.006)	(0.009)	(0.007)	(0.006)	(0.008)	(0.006)
p-value difference	0.661	p<0.001	0.165	0.133	0.001	0.320	p<0.001
Observations	12,532	31,665	7,247	26,764	24,680	22,977	28,467
				Panel B: Test scores			
September effect for boys	0.210***	0.196***	0.159***	0.207***	0.180***	0.205***	0.186***
	(0.013)	(0.008)	(0.015)	(0.009)	(0.008)	(0.009)	(0.008)
September effect for girls	0.200***	0.200***	0.229***	0.195***	0.213***	0.205***	0.196***
	(0.011)	(0.007)	(0.014)	(0.008)	(0.008)	(0.008)	(0.008)
p-value difference	0.544	0.694	0.001	0.287	0.004	0.959	0.362
Observations	172,587	447,011	111,077	368,859	361,816	338,780	391,895
Number of individuals	33,132	84,946	21,133	70,701	68,510	64,342	74,869
				Panel C: Redshirted			
September effect for boys	-0.030***	-0.063***	-0.169***	-0.033***	-0.110***	-0.020***	-0.115***
	(0.002)	(0.002)	(0.005)	(0.002)	(0.002)	(0.001)	(0.002)
September effect for girls	-0.022***	-0.022***	-0.057***	-0.018***	-0.037***	-0.012***	-0.041***
	(0.002)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.002)
p-value difference	0.003	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001
Observations	33,132	84,946	21,133	70,701	68,510	64,342	74,869
				Panel D: Retained			
September effect for boys	-0.214***	-0.175***	-0.096***	-0.205***	-0.139***	-0.165***	-0.178***
	(0.007)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)	(0.004)
September effect for girls	-0.203***	-0.121***	-0.055***	-0.169***	-0.090***	-0.140***	-0.121***
	(0.007)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)
p-value difference	0.252	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001
Observations	33,132	84,946	21,133	70,701	68,510	64,342	74,869

Note: Sample is based on all singleton births between 1994 and 2000. For each sample and outcome we present two estimates on being born in September separately for males and females. The p-values reported below each estimates pair test statistical equality of the two coefficients. Columns (1) to (3) present heterogeneity by maternal education, columns (4) and (5) present heterogeneity by medicaid status which is proxy for income, and columns (6) and (7) present heterogeneity by race and ethnicity. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). Kindergarten readiness excludes cohorts 1997 to 1999 due to missing data. Redshirting is defined as indicator variable that equals to one if a child has a higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Control variables include: cohort indicators, log birth weight, gestational age, indicator for starting prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health problems. Heteroskedasticity robust standard errors for being redshirted and retained while clustered at individual level for test scores.

Table 5: Effects of school starting age (August vs. September) on juvenile delinquency and college outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
		Race/Ethnicity			SES	
	All	White	Minority	Low	Medium	High
	Panel A: Ever attended college					
September birth	0.013*	0.016**	-0.008	0.003	0.026**	0.010
	(0.007)	(0.008)	(0.016)	(0.013)	(0.013)	(0.012)
Mean of Y	0.609	0.638	0.497	0.474	0.619	0.739
	Panel B: Obtained college degree					
September birth	0.011*	0.012	-0.000	0.009	0.024**	0.001
	(0.007)	(0.008)	(0.013)	(0.010)	(0.012)	(0.013)
Mean of Y	0.330	0.365	0.188	0.186	0.323	0.484
	Panel C: Obtained college degree from competitive/selective institution					
September birth	0.013**	0.011	0.014	0.006	0.026**	0.024**
	(0.006)	(0.007)	(0.009)	(0.007)	(0.010)	(0.012)
Mean of Y	0.180	0.202	0.084	0.073	0.159	0.310
	Panel D: In juvenile justice system (indicator x 100)					
September birth	-0.149	-0.262*	0.113	-0.222	-0.129	-0.077
	(0.138)	(0.142)	(0.397)	(0.317)	(0.213)	(0.144)
Mean of Y	0.969	0.728	1.843	1.888	0.785	0.214
Observations	16,923	12,581	3,704	5,398	5,681	5,844

Note: Sample is based on all singleton births from one unnamed county in Florida for years 1981 to 1990. Outcome variables are: ever attending college (Panel A); graduating with a degree (Panel B); graduating with a degree from a competitive/selective institution defined as those rated "competitive" or above as recorded in the NCES-Barron's Admissions Competitive Index Data Files for 2004 (Panel C); and being in juvenile detention center by age 16 (Panel D). Column (1) presents estimates for all students, column (2) for white students, column (3) for African-American or Hispanic students. Columns (4) to (6) present results by terciles for neighborhood SES. All regressions are based on September vs. August comparison and OLS with following indicator variables as controls: limited English proficiency, ever receiving subsidized meal in school, ever living in low-income neighborhood. Heteroskedasticity robust standard errors.

Table 6: Effects of school starting age (August vs. September) - utilizing regional variation in scores, redshirting and retentions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
					Heterogeneity			
	Full sample		Maternal education		Medicaid birth		Minority	
		HS dropout	HS grad	College grad	Yes	No	Yes	No
				Panel A: Redshirting analysis				
September birth*% redshirted	-0.225*	-0.405*	-0.265	-0.159	-0.269*	-0.139	-0.035	-0.073
	(0.114)	(0.238)	(0.262)	(0.324)	(0.160)	(0.208)	(0.144)	(0.114)
% redshirted	0.078***	0.084	0.061	0.071***	0.028	0.054*	-0.076	0.061*
	(0.027)	(0.062)	(0.042)	(0.024)	(0.055)	(0.028)	(0.052)	(0.033)
September birth	0.699***	0.275***	0.315***	0.384***	0.414***	0.489***	0.483***	0.502***
	(0.069)	(0.082)	(0.097)	(0.128)	(0.051)	(0.091)	(0.046)	(0.050)
				Panel B: Retention analysis				
September birth*% retained	-0.188***	-0.100	-0.174**	-0.234	-0.135*	-0.299***	-0.121***	-0.243***
	(0.070)	(0.068)	(0.080)	(0.149)	(0.070)	(0.085)	(0.035)	(0.080)
% retained	-0.183***	-0.104*	-0.049	-0.294***	-0.111**	-0.145**	-0.141***	-0.180***
	(0.058)	(0.060)	(0.043)	(0.092)	(0.053)	(0.058)	(0.043)	(0.057)
September birth	0.433***	0.303***	0.258***	0.069	0.373***	0.164***	0.349***	0.142***
	(0.068)	(0.075)	(0.045)	(0.117)	(0.064)	(0.056)	(0.044)	(0.050)
N	910		449		712		624	
N (districts)	65		41		62		54	

Note: Sample is based on all singleton births between 1994 and 2000. All regressions are run on aggregated means data where the variables are collapsed at year of birth-school-district-September births level. Analysis is based on 65 school districts and we have to exclude two small school districts because they do not record place of residence at birth for years 1994 and 1995. These two districts constitute 1.5% of the full population of births in years 1996 to 2000. Each regression controls for school district and cohort fixed effects and weights the estimates by number of children in cohort-district-September birth cells. Panel A presents results for the relationship between test scores and redshirting while Panel B between test scores and early retention. Each panel presents estimates on indicator for September births, fraction of children experiencing given explanatory variable (redshirting or retention) and the interaction between these two variables. Column (1) presents results for full sample: 65 districts, 7 years and 2 birth months. Columns (2) to (4) present the analysis split additionally by maternal education (cell is district, year of birth, September birth and three maternal education groups). Columns (5) and (6) present the analysis split by maternal medicaid status (cell is district, year of birth, September birth and two maternal education groups). Columns (7) and (8) present the analysis split by racial/ethnic minority status (cell is district, year of birth, September birth and two maternal race/ethnicity groups). In the heterogeneity analysis we require at least five observations in each cell and all three/two heterogeneity dimensions in each cell. This yields unbalanced repeated cross-section in the heterogeneity analyses.Standard errors are clustered at school district level.

Appendix

Table A1: Descriptive statistics: demographic characteristics of mothers and children

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		August and September births						
	All births	All	Singletons sample used in analysis			Sibling sample used in analysis		
			All	August	September	All	August	September
% African-American	21.9	22.4	25.8	25.7	25.9	24.2	24.2	24.2
% Hispanic	22.7	23.0	24.1	23.9	24.3	23.6	23.6	23.6
% immigrant	23.0	23.3	23.1	22.8	23.4	20.0	20.0	20.0
% HS dropout	20.6	20.6	23.8	23.6	24.0	25.0	25.0	25.0
% HS grad	59.0	59.3	61.0	61.0	61.0	55.1	55.1	55.1
% college grad	20.5	20.1	15.2	15.4	15.0	19.9	19.9	19.9
% married	65.6	65.2	60.7	61.0	60.3	63.4	63.1	63.7
% Medicaid birth	44.4	45.1	50.8	50.5	51.1	50.4	50.4	50.4
% male	51.2	51.1	50.6	50.5	50.8	51.6	51.5	51.6
% mom health problems	23.7	23.7	24.3	24.4	24.3	23.2	23.0	23.4
Maternal age	27.1	27.1	26.6	26.6	26.6	24.8	24.8	24.8
Birth weight	3343	3341	3328	3325	3330	3318	3318	3319
% September	8.8	50.0	48.8	0.0	100.0	50.0	0.0	100.0
N	1,220,803	215,971	139,211	71,214	67,997	2,184	1,092	1,092

Note: Sample is based on all singleton births between 1994 and 2000. Table A1 present means and sample sizes for eight different samples. Column (1) includes all births between 1994 and 2000 with complete demographic information; column (2) presents a subset of these births from August and September. Columns (3) to (5) present information for children used in the singletons empirical analysis while columns (6) to (8) are restricted to sample of siblings used in the sibling fixed effects empirical analysis. Columns (3) and (6) present descriptives for pooled August and September births while columns (4), (5), (7) and (8) present it for each month and sample separately.

Table A2: Effects of school starting age (August vs. September) - separate estimates for mathematics and reading

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Grade 3 to 8 pooled test scores in math			Grade 3 to 8 pooled test scores in reading		
			Panel A: Singletons (OLS)			
September birth	0.186***	0.184***	0.190***	0.208***	0.207***	0.213***
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)
Observations		722,642			728,913	
Number of children		139,038			139,188	
			Panel B: Siblings (OLS)			
September birth	0.195***	0.192***	0.196***	0.239***	0.235***	0.238***
	(0.026)	(0.028)	(0.027)	(0.027)	(0.028)	(0.027)
Observations		10,758			10,874	
Number of sibling pairs		1,092			1,092	
			Panel C: Siblings (FE)			
September birth	0.195***	0.199***	N/A	0.239***	0.238***	N/A
	(0.026)	(0.026)		(0.027)	(0.026)	
Observations	10,758			10,874		
Number of sibling pairs	1,092			1,092		
			Panel D: Siblings with same parents (FE)			
September birth	0.209***	0.206***	N/A	0.237***	0.236***	N/A
	(0.031)	(0.031)		(0.031)	(0.031)	
Observations	7,392			7,456		
Number of sibling pairs	735			735		
Year dummies		X	X		X	X
Early health controls		X	X		X	X
Demographic controls			X			X

Note: Sample is based on all singleton births between 1994 and 2000. All estimates come from August vs. September comparison. Outcomes are grades 3 to 8 mathematics test scores in columns (1) to (3) and grades 3 to 8 reading test scores in columns (4) to (6). Samples are: universe of singletons (Panel A); siblings born one in each month (Panels B and C) and siblings born one in each month where the father is know and the same across the two births (Panel D). OLS regressions in Panels A and B while sibling fixed effects regressions in Panels C and D. Columns (1) and (4) do not include any controls; columns (2) and (5) control for cohort dummies, log birth weight, gestational age, indicator for start of prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health at birth; columns (3) and (6) further control for maternal education, medicaid status at birth, race, ethnicity and gender. Standard errors clustered at individual level.

Table A3: Effects of school starting age (August vs. September) - selection into public schools

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			Singletons				Sibling FE	
	P(matched to public schools)			P(observed with 3rd grade test scores | matched to public schools)				
September birth	-0.019***	-0.019***	-0.020***	0.006***	0.006***	0.005***	0.005	0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.010)	(0.010)
Mean of Y		0.807			0.818		0.833	
Observations		215,971			174,439		2,952	
Year dummies		X	X		X	X		X
Early health controls		X	X		X	X		X
Demographic controls			X			X		

Note: Sample is based on all singleton births between 1994 and 2000. All estimates come from August vs. September comparison. The dependent variable in columns 1 to 3 is probability of being matched between birth records and school records. The dependent variable in columns 4 to 8 is probability of being observed with third grade test score conditional on being matched between birth and public school records. Samples are: universe of singleton births (columns 1 to 3); universe of singleton births matched to public school records (columns 4 to 6); and subsample of siblings born one in each month (columns 7 and 8). Cross-sectional regressions in columns 1 to 6 and sibling fixed effects regressions in columns 7 and 8. Columns 1, 4 and 7 do not include any controls; columns 2, 5 and 8 control for cohort dummies, log birth weight, gestational age, indicator for start of prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health at birth (column 8 additionally controls for gender); columns 3 and 6 further control for maternal education, Medicaid status at birth, race, ethnicity and gender. Robust standard errors in columns 1 to 6 and clustered at family level in columns 7 and 8.

Table A4: Effects of school starting age (August vs. September) - heterogeneity by demographic characteristics of mother and child

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	All	Maternal education			Income		Race/Ethnicity		Gender	
VARIABLES		HS dropout	HS grad	College grad	Medicaid	Non-medicaid	Black/Hispanic	White	Male	Female
					Panel A: Kindergarten readiness					
September birth	0.100***	0.150***	0.096***	0.043***	0.137***	0.065***	0.132***	0.079***	0.111***	0.090***
	(0.004)	(0.009)	(0.005)	(0.007)	(0.006)	(0.004)	(0.006)	(0.004)	(0.005)	(0.005)
Observations	51,444	12,532	31,665	7,247	26,764	24,680	22,977	28,467	25,861	25,583
					Panel B: Test scores					
September birth	0.195***	0.205***	0.198***	0.192***	0.201***	0.196***	0.205***	0.191***	0.194***	0.198***
	(0.004)	(0.008)	(0.005)	(0.010)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	730,675	172,587	447,011	111,077	368,859	361,816	338,780	391,895	367,468	363,207
Number of individuals	139,211	33,132	84,946	21,133	70,701	68,510	64,342	74,869	70,504	68,707
					Panel C: Redshirted					
September birth	-0.050***	-0.026***	-0.043***	-0.114***	-0.026***	-0.074***	-0.016***	-0.078***	-0.072***	-0.027***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	139,211	33,132	84,946	21,133	70,701	68,510	64,342	74,869	70,504	68,707
					Panel D: Retained					
September birth	-0.151***	-0.209***	-0.148***	-0.076***	-0.187***	-0.115***	-0.152***	-0.150***	-0.172***	-0.129***
	(0.002)	(0.005)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	139,211	33,132	84,946	21,133	70,701	68,510	64,342	74,869	70,504	68,707

Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). Column (1) presents average estimates; columns (2) to (4) present heterogeneity by maternal education, columns (5) and (6) present heterogeneity by medicaid status which is proxy for income; columns (7) and (8) present heterogeneity by race and ethnicity where minority is defined as either African-American or Hispanic, and columns (9) and (10) present heterogeneity by gender. Kindergarten readiness excludes cohorts 1997 to 1999 due to missing data. Redshirtingk is defined as indicator variable that equals to one if a child has a higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Control variables include: cohort indicators, log birth weight, gestational age, indicator for starting prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health problems. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Table A5: Effects of school starting age (August vs. September) - heterogeneity by deciles of birth weight

	Deciles of birth weight									
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Kindergarten readiness									
September birth	0.142***	0.113***	0.098***	0.096***	0.085***	0.095***	0.088***	0.100***	0.100***	0.088***
	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
Observations	5,294	4,823	5,496	5,490	4,849	4,848	5,455	5,408	4,847	4,934
	Panel B: Test scores									
September birth	0.187***	0.216***	0.181***	0.195***	0.185***	0.198***	0.198***	0.201***	0.189***	0.206***
	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.013)	(0.013)	(0.015)	(0.014)
Observations	75,864	69,780	77,920	78,984	68,510	68,914	77,602	76,677	63,149	73,275
Number of individuals	14,580	13,311	14,846	15,011	13,045	13,132	14,727	14,621	12,020	13,918
	Panel C: Redshirted									
September birth	-0.045***	-0.044***	-0.043***	-0.043***	-0.045***	-0.047***	-0.053***	-0.054***	-0.060***	-0.064***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	14,580	13,311	14,846	15,011	13,045	13,132	14,727	14,621	12,020	13,918
	Panel D: Retained									
September birth	-0.181***	-0.161***	-0.156***	-0.153***	-0.144***	-0.144***	-0.151***	-0.139***	-0.139***	-0.134***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.007)	(0.006)
Observations	14,580	13,311	14,846	15,011	13,045	13,132	14,727	14,621	12,020	13,918

Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison for each decile of birth weight. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). Kindergarten readiness excludes cohorts 1997 to 1999 due to missing data. Redshirting is defined as indicator variable that equals to one if a child has a higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Control variables include: cohort indicators, indicator for starting prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health problems. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Table A6: Effects of school starting age (August vs. September) - heterogeneity by gestational age

	(1)	(2)	(3)	(4)	(5)	(6)
	Gestational length					
VARIABLES	Very preterm	Preterm	Early term	Full term	Late term	Post term
	Panel A: Kindergarten readiness					
September birth	0.172***	0.123***	0.101***	0.093***	0.102***	0.182***
	(0.043)	(0.016)	(0.008)	(0.005)	(0.011)	(0.026)
Observations	565	3,292	11,551	29,801	5,155	1,080
	Panel B: Test scores					
September birth	0.262***	0.177***	0.200***	0.195***	0.193***	0.238***
	(0.044)	(0.017)	(0.009)	(0.006)	(0.014)	(0.034)
Observations	7,839	49,218	175,530	416,919	69,029	12,140
Number of individuals	1,528	9,444	33,720	79,181	13,058	2,280
	Panel C: Redshirted					
September birth	-0.068***	-0.050***	-0.047***	-0.050***	-0.053***	-0.039***
	(0.010)	(0.003)	(0.002)	(0.001)	(0.003)	(0.006)
Observations	1,528	9,444	33,720	79,181	13,058	2,280
	Panel D: Retained					
September birth	-0.208***	-0.166***	-0.158***	-0.147***	-0.139***	-0.150***
	(0.023)	(0.009)	(0.004)	(0.003)	(0.007)	(0.017)
Observations	1,528	9,444	33,720	79,181	13,058	2,280

Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison for each gestational age group. Gestational age groups are defined as follows: very preterm - below 32 weeks, preterm - 32 to 36 weeks, early term - 37 to 38 weeks, full term - 39 to 40 weeks, late term - 41 weeks, and post term - above 41 weeks. Outcomes are: kindergarten readiness (Panel A), pooled math and reading test scores in grades 3 to 8 (Panel B), probability of being redshirted (Panel C) and school retention (Panel D). Kindergarten readiness excludes cohorts 1997 to 1999 due to missing data. Redshirting is defined as indicator variable that equals to one if a child has a higher than expected, based on date of birth, age at the time of first observation in school records in either kindergarten or grade one. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Control variables include: cohort indicators, indicator for starting prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health problems. Heteroskedasticity robust standard errors for kindergarten readiness, being redshirted and retained while clustered at individual level for test scores.

Table A7: Effects of school starting age (August vs. September) - heterogeneity by school quality

	Deciles of school quality									
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Test scores									
September birth	0.207***	0.165***	0.196***	0.213***	0.195***	0.178***	0.204***	0.190***	0.193***	0.198***
	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)
Observations	69,241	69,901	69,854	69,636	70,291	69,699	70,220	70,078	70,341	70,576
Number of individuals	13,259	13,342	13,296	13,251	13,365	13,233	13,358	13,305	13,335	13,345
	Panel B: Retained									
September birth	-0.161***	-0.162***	-0.162***	-0.161***	-0.149***	-0.155***	-0.153***	-0.145***	-0.130***	-0.124***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.005)
Observations	13,259	13,342	13,296	13,251	13,365	13,233	13,358	13,305	13,335	13,345

Note: Sample is based on all singleton births between 1994 and 2000. Each point estimate reflects August vs. September comparison for each decile of contemporaneous school quality. Outcomes are: pooled math and reading test scores in grades 3 to 8 (Panel A); and and school retention (Panel B). For definition of school quality see Section 2.1. School retention prior to grade three is defined as an indicator variable that equals to one if child is observed twice in the same grade. Control variables include: cohort indicators, log birth weight, gestational age, indicator for starting prenatal care in first trimester as well as indicators for congenital anomalies, abnormal conditions at birth and maternal health problems. Heteroskedasticity robust standard errors for being retained while clustered at individual level for test scores.