What a Difference a Day Makes: Estimating Daily Learning Gains During Kindergarten and First Grade Using a Natural Experiment

Maria D. Fitzpatrick
SIEPR, Stanford University

David Grissmer
CASTL, University of Virginia

Sarah Hastedt
University of Virginia

Abstract
Knowing whether time spent in formal schooling increases student achievement, and by how much, is important for policymakers interested in determining efficient use of resources. Using the ECLS-K, we exploit quasi-randomness in the timing of assessment dates to examine this question. Conservative estimates suggest a year of school results in gains of about one standard deviation above normal developmental gains in both reading and math test scores. The results are statistically significant and extremely robust to specification choice, supporting quasi-randomness of test dates. Estimates of skill accumulation due to formal schooling do not vary based on socioeconomic characteristics.

Keywords: School Resources, Natural Experiment, School Year Length
JEL Classifications: I21, I28

Corresponding Author: Maria Fitzpatrick, Stanford Institute for Economic Policy Research, mfitzpat@stanford.edu.

1. Introduction

There has been a large body of literature devoted to estimating the returns to public investments in educational resources for children. This is because improving child outcomes has the potential to benefit society through increased economic growth and decreased dependence on government-funded social insurance programs. Educational interventions can be classified as taking one of two forms, those that improve the quality of the inputs and those that increase their quantity. Much of the work to date has focused on measuring the effects of interventions aimed at increased quality, such as teacher quality, class size or classroom climate (Hanushek 1998, 2003; Krueger 2003) or on estimating the effects of all school inputs, including those on both the quantity and quality margins (Card and Krueger 1992; Loeb and Bound 1996). However, investments in resources are likely to happen concurrently – historically, school districts with the longest term lengths were those with the highest paid teachers – making it difficult to disentangle the effects of changes in quality from those in quantity. This work joins a more recent strand of the literature focused on determining the returns to increasing the quantity of inputs, namely time in school, separately from changing their quality (Hansen 2008; Marcotte and Hemelt 2008; Marcotte 2007; Pishke 2007; Simms 2008; Leuven et al. forthcoming).

We estimate the effect of additional time in school on children's math and reading scores early in their schooling careers. Because the natural experiment we employ involves quasi-randomization of the time between tests, we are not plagued by the correlation between our measure of term length and changes in measures of quality. Further, because we use variation in the time between tests that occurs over several months, we provide a more comprehensive picture of the effects of schooling on achievement than previously seen. By investigating the effects of extra time in school on children's academic achievement, we hope to help answer the question of how efficient public funds spent on schooling are. If the daily rate of learning is constant over the course of the year (a fact for which we provide evidence below), the extent to which there are increases in academic achievement resulting from another day of school at the margin may also help to inform policy debates about the returns to lengthening the school-year.

To measure how much children learn per day of schooling we exploit variation in the timing of test taking in the Early Childhood Longitudinal Study – Kindergarten Class of 1998 (ECLS-K). Because children in the study were not all assessed on the same day, there is variation in the number of days elapsing between tests. The timing of tests appears to be essentially random (henceforth it will be referred to as quasi-random), allowing us to use the variation in the number of days between tests to measure the average effect of an additional day between tests. Children's math and reading scores increase by about 1.5 standard deviations with the passage of a school year.

We then isolate the effect of an additional school day in spring or fall separately from an additional day of development without the expense of school resources by first comparing the patterns of gains across the school-year and over the summer and then by using the within-school variation in ages at test taking time that is driven by date of birth. The results suggest that about two-thirds of the gain over the school-year-period (or one standard deviation) is due to time spent in school. Finally, we examine the effects of an extra day of school between assessments across different subgroups of the population based on gender, race, parental education and family income and find surprisingly little variation.

The remainder of the paper is laid out as follows. Section 2 summarizes recent studies of the relationship between time spent in school and academic achievement and describes the ECLS-K data used in our analysis. In section 3, we discuss estimation and our identification strategy. We then present results robust to specification choice that suggest there are gains associated with additional time in school. Section 4 discusses heterogeneous effects and Section 5 concludes.

2. Background on the Effects of Time in School and the Data

2a. Effects of Time in School

Previous studies examining the effect of time spent in school focus on the impact of variation in term length on academic performance and find somewhat mixed results. Pischke (2007) exploits variation in instructional time created by the German “short school years” of 1966 and 1967 and finds that the shorter school years were associated with a small increase in grade repetition in primary school and a decrease in enrollment in more advanced secondary school tracks. Lee and Barro (2001) use international data and find that the length of a school term is positive and significantly associated with math and science test scores and negative and significantly associated with reading test scores. Term length is negatively related to grade repetition rates and drop-out rates but the latter finding is not statistically significant.

Simms (2008) looks at variation in class time prior to a standardized test while the overall length of the school year was unchanged. Variation in time prior to testing was generated by the implementation of a law in Wisconsin restricting school districts to start dates after September 1st. He finds that additional class time is associated with small increases in math scores among fourth grade students but is not associated with reading score gains. Extra classroom days are also associated with improved reading scores among third grade students in the upper end of the ability distribution.

In work similar in spirit to Simms (2008), Marcotte and Hemelt (2008), Marcotte (2007) and Hansen (2008) exploit changes in the number of school days missed due to inclement weather. Hansen (2008) supplements the analysis of weather-related closings with an investigation of the effects of state-mandated test date changes in Minnesota. All find that more time in school before tests improves student performance on state-wide exams.

Finally, there is a strand of literature using the variation in age and amount of schooling based on eligibility cutoff dates (e.g. Cahan and Davis, 1987; Cahan and Cohen, 1989; Luyten, 2006, Neal and Johnson, 1996; Gormley and Gayer, 2005; Cascio and Lewis, 2006). These studies generally find that time in school improves academic outcomes of children, particularly minorities. For example, Leuven et al. (forthcoming) exploit variation in the amount of schooling young children receive in Holland to estimate the effect of time in school on test scores. The variation in the amount of schooling is driven by policies that allow children to start school immediately after they turn four and place children born around summer in the same class. These age-related rules create variation in the amount of schooling by up to 11 weeks and the study suggests this increases language and math scores of disadvantaged students by about 0.05 standard deviations.

Our main contributions to the literature on the effects of time spent in school are twofold. First, as we show in evidence below, our identification strategy is both novel and credible. Second, but somewhat related, the differential exposure to school is continuous over a large range of days between tests: some students had as few as 60 days between tests while others had over 200. This allows us to estimate the returns to time in school over a longer time between tests and investigate whether rates of learning vary over the course of the year. By measuring the effect of time in school over the course of the "normal" school year, we complement the existing literature which has largely only been able to use small variations in class time due to weather shocks or changes in school start dates.

2b. Data: Early Childhood Longitudinal Study-Kindergarten Class of 1998-1999

We use data from the ECLS-K. The study began collecting data on kindergarteners in the 1998-1999 school-year. Because the National Center for Education Statistics (NCES) designed the study in part to determine what effects kindergarten experiences had on children, students were tested in the fall and spring of their kindergarten year. The same test materials were used for both assessments. These assessments consisted of questions in three subject areas: language and literacy (reading), mathematical thinking (math) and “knowledge of the social and physical world” (general knowledge).[1] Administrators evaluated children in one-on-one settings by asking them to respond to questions orally or by pointing to the answer. The testing sessions generally lasted about 45 minutes. The language and literacy assessment measured basic skills, such as letter recognition and knowledge of beginning and ending sounds, as well as vocabulary and listening comprehension. The mathematical skills measured included

[1] ECLS-K Base Year User’s Manual (2001). Because understanding of what the general knowledge test measures is vague, we use only the reading and mathematics test scores.

identifying one-digit numbers, counting, recognizing geometric shapes and performing simple arithmetic and multiplication problems.

A two-stage testing procedure was used in the ECLS-K assessments. All children received the same set of first round questions, called routing questions. A child's performance on these routing questions determined the difficulty level of the questions the assessor asked the child in the next round. The pattern of a child's responses to the routing and second stage questions was used to develop an IRT scale score, which is comparable across children.[2] We use these IRT scale scores in our analysis.

The base year sample of the ECLS-K includes 21,700 observations.[3] Children who are repeating kindergarten may learn at different rates than those who are first-time kindergarteners; therefore we drop the 680 students in the sample who are in kindergarten for the second time. An additional 4,880 observations are deleted from the sample because they represent children who did not have either a fall or spring test score in at least one of the two subjects. Our main sample includes 16,150 students.

We argue that the natural experiment is random and so, like in a clincal trial, if the randomization is done correctly we should not need to control for observable characteristics. To verify the robustness of our results we conduct several analyses that incorporate additional information about the children beyond their test scores and the time elapsing between assessments. For many children, information on socioeconomic

[2] Item response theory (IRT) uses the pattern of right, wrong, and omitted responses among items included in the test actually administered to students in combination with information on the difficulty and "guess-ability" of each test item to place students on a continuous ability scale. This allows for comparison of scores between students regardless of which second-stage test form was actually taken (ECLS-K Base Year User's Manual, 2001).

[3] In accordance with confidentiality policy of NCES, sample sizes have been rounded to the nearest 10 observations.

characteristics is missing. When we include these characteristics in the estimation we use dummy variables to control for observations with missing data for some variables.[4]

Means and standard deviations of the observable characteristics of our full sample of kindergarteners are in the first and second columns of Table 1, respectively. Test scores have been normalized using the mean and standard deviation of the fall kindergarten test scores for children in our sample. The average gains in scale scores from the fall to spring tests in reading and math are 1.18 and 1.13 standard deviations, respectively. However, the difference between fall and spring test scores ranges widely with some students gaining considerably more than the average and others losing a little ground between the fall and the spring testing.[5]

In our sample, half of the children are boys. African-Americans comprise 16 percent of the sample and Hispanic children make up another 13 percent. Thirty percent of children in the sample have mothers with a high school education and another third have mothers who have taken some college courses. The remaining are split among those that have a BA (16 percent), those that have a graduate or professional degree (8 percent) and those that have not received a high school diploma (10 percent). Lastly, the children are fairly spread out between the regions of the country and among location types (urban, suburban, rural).

[4] This sample size is in line with the size of the sample used by most researchers employing the ECLS-K data (e.g. Elder and Lubotsky 2006, Fryer and Levitt 2004). Students for whom at least some data is missing are, on average, poorer and score lower on the tests they do take than those for whom all information is available. We replicated our analysis using several methods of correcting for missing data, including multiple imputation methods (see Royston 2006), limiting the sample to just those students with valid information and using information about other teachers and students in the school as proxies for the child's/teacher's information. Regardless of our choice of methods, the results were remarkably similar to those presented here.

[5] Not many children lose points over the school year and the great majority of children with negative gains between tests lose less than five points. Regression to the mean may explain part of these negative gains.

We also make use of a sub-sample of children tested in the fall of first grade. The NCES chose 30 percent of the original sample of schools for data collection in the fall of first grade. This sub-sample includes nearly 5,300 children and represents 27 percent of the children in the base year sample. The NCES also attempted to re-test the entire original sample in the spring of 2000, when most students were in their first grade year. The fall and spring first grade testing focused on the same three areas of competence as kindergarten testing. The same two-stage testing procedures were used, but the reading assessment administered in the fall of first grade was an augmented version of the kindergarten assessment.[6] Descriptive statistics for our first grade sub-samples, which mirror the kindergarten samples in terms of sample selection procedures, are reported in the second set of columns of Table 1. This subsample is quite similar in its characteristics to the entire ECLS-K sample.

3. Quasi-Random Variation in Test Dates and Estimation of Effects of School Days on Achievement

3a. Assignment of Test Dates

The assignment of test dates is critical to our estimation strategy. The interpretation of our estimates as the average effect of extra time in school between tests hinges upon our argument that the test dates were randomly determined, or at least were not correlated with any observable or unobservable characteristics that also affect gains in

[6] An NCES analysis of reading scores from kindergarten indicated that more students than expected scored near the test ceiling. The number of items on the reading assessment was therefore increased to include more difficult vocabulary and text to minimize the potential for ceiling effects. We have done our analysis using both the original scores and the scores as recalculated by NCES to account for the possible ceiling effects on the original assessments of kindergarteners. We present results using the original scores because our tests for ceiling effects (discussed in detail later) suggested there were none and the estimates using the original score are more conservative. Results using the recalculated scores are available from the authors upon request.

test scores. Ideally, children would have been randomly assigned to test dates in the fall and spring of kindergarten and first grade. Random assignment was not employed, but the process used to determine test dates produced what looks like a random distribution of test dates. However, the lack of strictly random assignment means that we will be unable to eliminate the possibility that test date assignment is correlated with unobservable characteristics determining test scores. The strongest argument in favor of our use of the natural experiment is that these were low stakes tests.[7]

This section examines the assignment process and its outcomes to provide significant evidence congruent with our argument that test dates have the characteristics of a random sample. A similar process was followed in all testing periods, so we focus our analysis on the kindergarten administrations. Panels A and B of Figure 1 show the wide distribution of test dates in the fall and spring, respectively.

The sequential pattern of test dates largely stemmed from the nature of the assessment used, which required one-on-one interaction between test administrators and the students in the more than 1,000 schools sampled. With such assessments, reliability is increased by limiting the number of test administrators. Hence the design of the ECLS-K was to have fewer administrators and spread the testing period more widely across the fall and spring schooling periods. The NCES preferred to survey students at the same point in time relative to their school start date, but the final decision about when to schedule the assessments was made by school administrators (in consultation with the affected teachers). Each administrator made an essentially independent decision.

[7] For a discussion and analysis of how performance on tests varies based on the "stakes" attached to them, see Jacob (2007).

The role of school administrators in scheduling assessments could be problematic for our identification strategy. Estimates of the effects of an extra day of kindergarten might be biased if teachers or administrators were scheduling children's assessments to maximize the gain between tests by attempting to put the most distance possible between the two tests. If so, Figure 1 would show the largest densities of observations clustering near the earliest possible test dates in the fall and the latest test dates in the spring; Figure 1 shows little indication of this type of pattern.

Another potential concern is that only the administrators and teachers who expect their children to benefit most from extra time between tests manage to "game" the test scheduling. Then, the administrators and teachers with the most to gain might schedule their tests with as much distance as possible between them, leaving the relatively narrow testing periods to other administrators and teachers (those with less to gain). If this type of behavior occurs when schedules are set, a regression estimating the spring test date as a function of the fall test date would return a negative coefficient. This could bias our estimates of the effect of a day between tests upward. However, this is not the case; the coefficient estimate is 0.161 with a standard deviation of 0.007. In fact, an NCES protocol stated a preference that students that were tested earlier (later) in the fall should be also tested earlier (later) in the spring. If this protocol had been followed to the letter, the aforementioned coefficient estimate would be one and there would be little variation in the number of days between tests, which is not the case.

Though the protocol does not seem to have been well-followed, to the extent that school starting dates are related to test dates, our estimates may be biased. School start dates are almost always set by school districts within state-determined “windows.”[8] If schools that start earlier (later), also show earlier (later) fall test dates, and starting date is correlated with test scores and days between tests or other observable or unobservable variables, then a potential bias issue also arises.

To address the concern that factors affecting achievement are also associated with test dates, we regress days between tests on a rich set of covariates including individual demographics and teacher characteristics (Table 2). Though some coefficients are statistically significant, only the relationship between being Hispanic and the number of days between tests is statistically significant across all three columns. What is more, only a very small portion of variance is explained (the R-squared statistics are all less than 0.07).

In order to get a sense for how close to random the assessment timing was, we conducted permutation tests which involved assigning each child a randomly drawn value of days between tests from a normal distribution with the same mean and standard deviation as the observed distribution of days reported in Table 1. We then regressed the randomly assigned number of days between tests on the observable characteristics in Table 2. We repeated the process 1,000 times and found that 12.7 percent of the coefficient estimates were statistically significant at the 5 percent level. This is less than the proportion of statistically significant estimates in Table 2 using the actual data (15 percent). This indicates that the number of days between assessments may be weakly correlated with observable characteristics, though the estimates in Table 2 do not provide

[8] These windows do vary by state (not before Labor Day, after Aug 20th, etc). Nationwide school starting dates range from Aug 15 to around Sept 10. These start dates can reflect, for instance, that students in some places can use the extra time before their school’s starting date to work (e.g. farming communities or beach towns).

any obvious patterns that this would bias our results.[9] To be sure, we conduct several robustness checks, incorporating variables corresponding to the observable characteristics in Table 2, along with state and school-start-week fixed effects in our equations, to check the sensitivity of results to their inclusion.

3b. Effect of an Additional Day between Tests in Kindergarten

We exploit the natural experiment by employing the variation in the number of days included in the period between which children are assessed. If test date assignment is essentially random with respect to test scores, our analysis allows us to measure the average marginal effect of extra time between tests. Adding a "typical" extra day between tests in our sample during the kindergarten year can be thought of as a weighted combination of adding a weekday/school-day (in which there are school and home inputs) and a weekend/nonschool-day (in which there are most likely only home inputs). In this combination, the weights on each type of day are the probability of the specific type of day occurring.[10] In this section, the question we attempt to answer is: *on average, what is the effect of extra time between tests (including both home and school inputs) on children's reading and math skills?*

---

[9] It is not the case, for example, that female students, who tend to score higher on reading tests, have more time between tests, on average, than their male counterparts. If this had been the case, we would have been concerned that the faster learning girls were driving the correlations between time elapsed between assessments and test scores. Further, the only coefficient that is statistically significant across the three separate estimations is that on the dummy variable for Hispanic ethnicity. The estimates suggest Hispanic students have less time between their assessments. If associated with the rate of learning at all (a conjecture we provide evidence refuting below), we hypothesize that Hispanic students do not learn as quickly as their Caucasian counterparts. Therefore the negative relationship between being Hispanic and the time between assessments biases our estimates of the effect of time on learning downward.

[10] If students went to school on every weekday and did not go to school on the weekends, the weights would be 5/7ths and 2/7ths, respectively. Because school schedules vary due to holidays, teacher work-days and other days on which school is not held (e.g. snow-days), the weight on school-days will actually be less than 5/7ths.

In our first attempt to measure the effect of time on skill accumulation, we use the variation in the number of days between the two tests to measure how much children are learning in a "typical" school year. Specifically, we estimate the equation

$$Gains_i = \alpha_1 SchoolYear_i + \varepsilon_i .^{11} \quad (1)$$

$Gains_i$ is the difference between the spring and fall test scores in a particular subject for child *i*. $SchoolYear_i$ represents the fraction of a school-year (180 school-days plus corresponding weekends, for 250 days) between the child's fall and spring assessments. We can therefore think of the specification as being related to a value added specification, where the change in test scores is related to the time that passes between tests; hence there is no constant term. The implicit assumption is that the relationship between an extra day between tests and the gains on those tests is linear. α_1 can be interpreted as the average marginal effect of a "typical" school-year-period on children's test scores.

Table 3 reports estimates of α_1 when the dependent variables are gains in reading (Panel A) and math (Panel B).[12] In the spirit of our argument that the test dates are random, column (1) reports results from regressions using the full sample and does not include any controls or fixed effects, much as analysis of a randomized trial might not include observed characteristics. The passage of a school-year between tests is associated with additional gains of 1.586 standard deviations in reading and 1.510

[11] Todd and Wolpin (2003) describe the assumptions implicit with various specifications of the education production function. In this specification, it is assumed that previous inputs only enter the production function through their effects on lagged test scores. Our results are not sensitive to our specification of the relationship between gains on test scores and fall test scores; results are available from the authors upon request.

[12] Because of sample selection related to missing information, we do not use ECLS-K sample weights, but our results do not change much at all if we do. Results using sample weights are available from the authors upon request.

standard deviations in math. With each of the dependent variables the estimated effect of an extra day between tests is statistically significant.

Column (2) and (3) report results when controls are added for child characteristics and school and teacher characteristics, respectively. The inclusion of these controls decreases the size of the estimates slightly to 1.164 standard deviations in reading and 1.283 standard deviations in math. The estimates for math when controls are included are not statistically different than those in the previous specifications.

If there are ceiling effects associated with these tests, the estimated relationship linking days between tests and test scores could be purely mechanical. That is, if the number of days spent in school before an assessment improves test scores, students taking the fall test later in the year will have higher test scores in the fall. With ceiling effects, these students' gains scores are held artificially low because there is less room for them to grow. This could lead to an estimated relationship between days and test scores partially driven by the assessment measure rather than only by the effects of time in school. Though IRT test scores are designed to eliminate ceiling effects, we investigate the possibility of their existence.

Koedel and Betts (2008) argue that a test score distribution's skewness is indicative of the presence of ceiling effects. Specifically, a negatively skewed distribution of test scores implies that there was bunching of children at the top end, suggesting ceiling effects. Figure 3 presents kernel density estimates of the distributions of fall and spring kindergarten reading and math scores. All are positively skewed, with estimates of the density skewness of at least 0.41.[13] Ceiling effects are not an issue with these assessments.

3c. Natural Development versus School Inputs

Our estimates suggest that the academic achievement of children improves with additional time between tests. However, without knowing more about children's experiences over the course of a "typical" additional school-year-period it is impossible to determine whether these gains are attributable to time spent at school or whether they reflect development that would have occurred without the expense of school resources. We now perform additional analyses to attempt to separate the effects of time spent in school from gains that would occur even in the absence of schooling.

3.c.i. Using the School-year versus the Summer

The NCES followed the ECLS-K sample for several years. The vast majority of children tested in kindergarten were tested again in the spring of their first grade year (2000). As detailed previously, a subsample was also tested in the fall of 1999, the first grade year. We can use this subsample to estimate the average daily gains in academic achievement over three periods. Consider our regression of the gains in reading scores on the fraction of a school year that elapses between the fall and spring kindergarten tests. The coefficient estimate represents the average marginal gain over the school-year period. This included schooldays and weekends and school holidays. The period between the fall assessment in first grade and the spring assessment in first grade includes

[13] Skewness is 1.58 for the fall reading test, 0.96 for the fall math test, 0.84 for the spring reading test and 0.41 for the spring math test.

a similar mix of schooldays and non-schooldays that is governed by the school schedule. The summer period, on the other hand, includes a much larger proportion of non-schooldays relative to the number of schooldays. We can therefore compare the estimates of gains over the school-year to those over the summer to help differentiate between developmental growth and time spent in school.

The results of regressions making use of the extra information from the first grade subsample are in Table 4. In the table, each set of rows represents a separate set of regressions where the dependent variable is gains in test scores in the subject indicated by the column header over the period indicated by the row header. All periods have been normalized to include 250 days, so that the coefficients of interest can be easily compared. The first thing to notice is that the gains in reading and math scores over the school-year-period of first grade are even larger than those over the kindergarten school-year-period. In reading, children gain an average of 2.436 standard deviations in test scores and, in math, children gain 1.716 standard deviations. Secondly, the coefficient estimates for the rate of learning over a 250 day period from the spring of kindergarten to the fall of first grade are much smaller than the school year estimates. For reading and math, the estimates imply that the rate of learning over the summer is at most about three-fourths of the rate of learning during the year.

The statistically significantly smaller coefficient estimates for the summer imply that the average rate of learning per day for both math and reading is slower in the “summer” period between tests than it is over the "school-year" periods. The majority of children likely experience little schooling over the summer period. In fact, in the average school-year-period there are two school-days per non-school-day, while in the average summer period the ratio is reversed.[14] The slower rate of learning over the summer period suggests that our estimates of the positive effects of time between tests are largely due to children's school participation in the fall and spring. This is consistent with the findings that children learn more and learn more efficiently during the part of the year when they are in school relative to the summer months when little schooling occurs (Alexander, Entwisle, and Olson 2001, Cooper et al. 1996).

3.c.ii. Using Within School Variation in Age

The second way we distinguish the effects of natural development from those of school resources is to use the within-school variation in age at testing for all children taking the test on the same day. More specifically, because all children within a school start school on the same day, those that also are assessed on the same day will have had the same number of school-days prior to assessment. Because there is variation in when children are allowed to start school (due to eligibility cutoffs) and when parents decide to send their children to school, children assessed on the same day will differ in age. We can use the difference between the effect of an extra 250 days during the schooling period and the effects of an extra 250 days in age at the time of testing to determine the effects of time in school itself.

To be clear, we estimate

$$SpringScore_i = \beta_0 + \beta_1 SchoolYear_i + FallScore_i + \omega_i \qquad (2)$$

[14] A back of the hand calculation equating the sum of the number of school-days and non-school-days in each period to the coefficient estimates for each period suggests that the rate of learning per non-school-day is less than 10 percent of the rate of learning per school-day. Because we do not have school schedules, we cannot precisely measure the amount of school-days versus non-school-days and so report this exercise as only suggestive evidence of the underlying relationship. Results are available from the authors upon request.

to determine the effects of a school year, including natural development and time in school, on children's gain in test scores. To isolate the effect of just development, we estimate

$$SpringScore_i = \gamma_0 + \gamma_1 Age250_i + \sum_j \sum_k \gamma_{jk} Startdate_j \times Testdate_k + \mu_i \quad . \qquad (3)$$

In this equation, there are fixed effects for school-starting-date by spring-test-date ($Startdate_j \times Testdate_k$). $Age250$ measures the number of 250 day periods in a child's life that have taken place at the time the child is assessed in the spring. The estimated coefficient of interest, γ_1, therefore measures the effects of being 250 days older on test scores.

The results of this exercise are presented in Table 5. Columns (1) and (4) present the results of estimating equation (2) with reading and math scores of kindergarteners in the spring as the dependent variables, respectively. The estimates suggest average gains of 1.543 and 1.412 standard deviations in reading and math, respectively, over the course of the school-year-period. The estimates of γ_1 from equation 3 are in columns (2) and (5). These estimates suggest that an additional 250 days in age contributes 0.34 and 0.51 standard deviations to reading and math scores, respectively. The difference between the estimates for each subject ($\beta_1 - \gamma_1$) can be used to approximate the gains attributable to schooling (under the assumption that the gains to development and to schooling are additively separable). Therefore, the results in Table 5 suggest that the gains in test scores attributable to schooling are 1.2 and 0.9 standard deviations in reading and math, respectively.

As mentioned earlier, a child's age at the time of the test is partly determined by age at school entry laws and partly by parental choice about when to enroll a child in kindergarten (e.g. red-shirting). Therefore, the results just discussed may confound the actual effects of age with other things related to the choice of when to enroll a child (e.g. family resources that can be devoted to keeping a child out of school until he is older may also improve test scores). In order to determine the effects of age alone, we follow Bedard and Dhuey (2006) by using birth date relative to the eligibility cutoff date for school entry. These results are in columns (3) and (6) of Table 5. Though the estimates are too noisy to be statistically significant, they are about the same size or smaller than the estimated effects of age obtained when not instrumenting. This supports our conclusion that much of the gain over the school-year-period is attributable to time in school itself.

4. Differential Treatment Effects

4a. Race, Gender, Poverty Status and Maternal Education

While we find evidence that an additional school-day between tests is associated with statistically significant gains in reading and math, the effect may vary across different student subgroups of the population. To test for heterogeneous effects, we use the parsimonious specification and include dummy variables for race and maternal education categories, poverty status, and interaction terms between these variables with each other and with our *SchoolYear* measure. None of the coefficient estimates on these characteristics were statistically significant, leaving us to conclude that the marginal gain to a typical day of schooling is not different for children in different population subgroups.

4b. Half- versus Full-Day Kindergarten

Though most states do not mandate the length of the school day for kindergarteners, many child advocates and policymakers call for full-day kindergarten to become the norm (for example, see the Education Commission of the States). The impact of an additional day between tests may depend on how many hours per day a child is in school. To explore this possibility, Table 6 presents results of regressions for children in half-day and full-day kindergarten separately. The estimated effects for a year of half-day kindergarten are slightly smaller than those for a year of full-day kindergartener. This may reflect the fact that although full-day kindergarten days are usually about twice as long as half-days, they do not devote twice as much time to instructional activity.[15]

That the estimates are not that different across full- and half-day kindergarteners might also suggest that the test score gains are attributable largely to development that occurs with age and not to differences in the amount of time spent in school. However, we are cautious about placing much emphasis on comparing differences in the estimates across the length of the kindergarten day as whether a child attends half-day or full-day kindergarten is likely correlated with unobservable characteristics that might also affect gains to schooling. This is not to say that our estimates do not accurately measure the effect of extra time in school on the population of full- or half-day kindergarteners separately. On the contrary, the same arguments for quasi-randomness in test days on the whole sample works for these subsamples, too. What is problematic, however, is comparison of the effects between the two subsamples because full- or part-day

[15] http://nces.ed.gov/pubs2004/2004078.pdf

kindergarten attendance is correlated with observable characteristics. [16] Although we find no evidence that the effects of time in school vary with observable characteristics, there may be heterogeneous effects across unobservable characteristics standing in the way of a valid comparison across these groups.

4.c. Nonlinearities in Effects

Underlying our estimation strategy has been the assumption that the rate of learning per day was constant over each period for which we estimate it. It is possible, however, that nonlinearities exist in the rate of learning across these periods. To investigate whether this is the case, we re-estimated equation (1) adding in higher order polynomials (quadratic and cubic) in the time between tests. We then conducted F-tests of the joint significance of these higher order terms and found we could not reject the null hypothesis that the coefficients are zero. Additionally, we also estimated models where where we divided the assessment periods into subgoups and estimated the effect of time between tests by subgroup (e.g. interacting the time between tests term with a set of dummies dividing the testing period into quintiles). These, too, failed to show any pattern of statistical significance suggesting nonlinearities. It appears that a constant rate of learning is an appropriate assumption over the periods we study.

[16] African American children and those living below poverty are more likely to be in full-day kindergarten while those living in suburbs or whose mothers have some college attendance but no degree are less likely to be in full-day kindergarten. We attempted to use state rules mandating full-day provision and policies determining state-funding levels for half-day classes as compared to full-day classes as instruments for the length of the kindergarten day a child attended. These attempts were unsuccessful because there was not enough cross-state variation in the rules and policies to also include state fixed effects.

5. Conclusion

This work adds to a growing body of literature examining the impact of instructional time on student performance. Consistent with previous work, our estimates suggest that there are test score gains for additional time spent in school. Conservative results show robust statistically significant estimates of gains to schooling over the course of the school-year of 1.2 standard deviations on reading tests and 0.9 standard deviations on math tests. If we convert these to gains per day of school, these estimates are smaller than those of Hansen (2008) and Marcotte and Hemelt (2008). One possible explanation for the difference in magnitude is that the aforementioned papers exploit variation in instructional time generated by unscheduled school closures due to inclement weather. These types of school closures may create significant disruption in instruction and other aspects of students' lives which may impact learning and test performance. In contrast, the variation we rely on does not have this disruptive effect. Importantly, the gains to extra time in school do not vary systematically across children based on observable characteristics or initial test scores

These results are important for helping researchers and practitioners understand how much children learn with an extra day of schooling. Our results suggest that there may be substantial positive effects on reading and math test scores if the school year were to be extended. To place our results in context, we can compare them to those of the famous class-size experiment, Tennessee STAR. In his cost-benefit analysis of STAR, Krueger (2003) hypothesizes that decreasing the class size from 22 to 15, on average, increases test scores by 0.020 in both reading and math by spending just 47% more per pupil. Were we to do a similar calculation and increase expenditure on school length by 47%, it would increase the number of school days by 84. Based on our estimates, this would result in an increase in test scores of 0.528 standard deviations in reading and 0.396 in math. Of course, this is under the assumption, (supported by our analyses) that the rate of learning is constant. Our results suggest increases in school quality on the extensive margin may have the potential to be just as effective as other targeted or untargeted intensive interventions. Our findings also have important implications for accountability policies. Researchers and policy-makers should take this per-day effect into account when studying and designing testing procedures, especially those tied to rewards or punishments for schools and/or teachers (Hansen and Marcotte 2006).

References

Bedard, K. and E. Dhuey (2006), The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects, Quarterly Journal of Economics, Vol. 121, pp. 1437-72.

Cahan, S., and N. Cohen (1989). "Age versus schooling effects on intelligence development." *Child Development*, Vol. 60, pp. 1239–49.

Cahan, S. and D. Davis (1987). "A between-grades-level approach to the investigation of the absolute effects of schooling on achievement." *American Educational Research Journal*, Vol. 24, pp. 1–12.

Card, D. and Krueger, A. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States, *Journal of Political Economy*, 100, 1-40.

Cascio, E. and E. Lewis (2006). "Schooling and the Armed Forces Qualifying Test. Evidence from School-Entry Laws." *Journal of Human Resources*, Vol. 41, pp. 294–318.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., and Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, 66, 227-268.

Elder, T.E. and Lubotsky, D.H. (2006). Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background, and Peers. Available at SSRN: http://ssrn.com/abstract=916533

Entwisle, D. R., Alexander, K. and Olson, L.S. (2001). Schools, Achievement, and Inequality: A Seasonal Perspective. *Educational Evaluation and Policy Analysis*, 23(2), 171-191.

Fryer, R. and Levitt, S. (2004). Understanding the Black-White Test Score Gap in the First Two Years of School. *The Review of Economics and Statistics*, 86(2), 447–464.

Gormley, W. and T. Gayer (2005). "Promoting School Readiness in Oklahoma. An Evaluation of Tulsa's Pre-K Program." *Journal of Human Resources*, Vol. 40, pp. 533–58.

Hansen, B. (2008). School Year Length and Student Performance: Quasi-Experimental Evidence. Mimeo. University of California – Santa Barbara.

Hansen B. and D. Marcotte (Forthcoming.) "Time for School: Instructional Time and Student Achievement." *Education Next.*

Hanushek, E.A. (2003). The Failure of Input-Based Schooling Policies, *Economic Journal*, Royal Economic Society, 113(485), F64-F98.

Hanushek, EA. (1998). The Evidence on Class Size, *Wallis Working Papers WP10*, University of Rochester - Wallis Institute of Political Economy.

Jacob, B. (2007). Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments. NBER Working Paper #12817.

Koedel, C. and Betts, J. (2008). Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. University of Missouri-Columbia WP 08-07.

Krueger, A.B. (2003). Economic Considerations and Class Size. *Economic Journal*, Royal Economic Society, 113(485), F34-F63.

Lee, J. and Barro, R. (2001). Schooling Quality in a Cross-Section of Countries. *Economica*, 68(272), 465-488.

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek and Dinand Webbink. (Forthcoming.) "Expanding Schooling Opportunities for Four Year Olds." *Economics of Education Review.*

Loeb, S. and Bound, J. (1996). The Effect of Measured School Inputs on Academic Achievement: Evidence from the 1920s, 1930s, and 1940s Birth Cohorts. *Review of Economics and Statistics*, LXXVIII, 653-664.

Luyten, H. (2006). “An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95.” *Oxford Review of Education*, Vol. 32, pp. 397–429.

Marcotte, D. and Hemelt, S. (2008). “Unscheduled School Closings and Student Performance.” *Education Finance and Policy*, 3(3), 316-338.

Marcotte, D. (2007). Schooling and Test Scores: A Mother-Natural Experiment. *Economics of Education Review*. 26(3): 629-640.

Neal D. and W. Johnson. (1996). “The Role of Pre-market Factors in Black-wage Differences.” *Journal of Political Economy*, Vol. 104, pp. 869–95.

Pischke, J. (2007). The impact of the length of school year on student performance and earnings: Evidence from the German short schooling years. *Economic Journal*, 117(523) 1216-1242.

Royston, P. (2006). ICE: Stata module for multiple imputation of missing values. Statistical Software Components S446602, Boston College Department of Economics, revised 25 Apr 2008.

Simms, D. (2008). Strategic responses to school accountability measures: It's all in the timing. *Economics of Education Review*, 27, 58-68.

Todd, P. and Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, 113, F3–F33.

U.S. Department of Education. (2001). *ECLS-K Base Year User's Manual*.

Table 1. Descriptive Statistics of the Characteristics of Children in the ECLS-K, Kindergarten Year

	Kindergarten Sample		1st Grade Sample	
	Mean	Std. Dev.	Mean	Std. Dev.
Scores				
Fall Reading	0.00	1.00	0.06	1.05
Fall Math	0.00	1.00	0.05	1.03
Spring Reading	1.18	1.22	1.27	1.25
Spring Math	1.13	1.20	1.19	1.22
Fall to Spring Gains in Scale Scores				
Reading	1.18	0.73	1.22	0.74
Math	1.13	0.69	1.15	0.70
Percent Male	0.50	0.50	0.50	0.50
Percent Black	0.16	0.37	0.15	0.36
Percent Hispanic	0.13	0.34	0.12	0.32
Percent of Mothers With:				
High School Diploma	0.29	0.45	0.28	0.45
Some College Attendance	0.32	0.47	0.33	0.47
Bachelor's Degree	0.16	0.37	0.16	0.37
Graduate or Professional Degree	0.08	0.27	0.09	0.28
Percent Living in:				
Urban Areas	0.40	0.49	0.38	0.48
Suburban Areas	0.37	0.48	0.36	0.48
Northeast	0.20	0.40	0.19	0.39
Midwest	0.27	0.44	0.27	0.45
South	0.32	0.47	0.31	0.46
Teacher Characteristics				
Age	39.52	14.08	39.82	13.65
Years of Experience	16.67	10.47	16.37	10.49
Percent Black	0.06	0.24	0.07	0.26
Percent Hispanic	.04	0.19	0.03	0.17
Percent with Some College	0.01	0.10	0.01	0.11
Percent with BA	0.58	0.49	0.58	0.49
Percent with Graduate Degree	0.33	0.47	0.32	0.46
Days Between Fall and Spring Tests	186.27	21.33	208.72	208.72
Number of Observations	16,150		4,020	

Notes: Based on the authors’ calculations using the ECLS-K without using sample weights.

Table 2. Relationship between Number of Days between Tests and Observable Characteristics

	Kindergarten (1)	**Summer K-1** (2)	**First Grade** (3)
Male	-0.001	-0.002	-0.003
	(0.001)	(0.004)	(0.003)
African American	-0.003	0.005	-0.014
	(0.005)	(0.013)	(0.009)
Hispanic	0.008*	-0.021*	0.017*
	(0.004)	(0.010)	(0.007)
Mother has HSD	-0.002	0.004	-0.002
	(0.003)	(0.010)	(0.005)
Mother has some college	0.000	0.004	-0.004
	(0.003)	(0.010)	(0.006)
Mother has BA	-0.004	0.001	-0.007
	(0.004)	(0.013)	(0.008)
Mother has graduate or professional degree	-0.006	0.001	-0.013
	(0.005)	(0.014)	(0.009)
Poverty	0.001	-0.011	0.011**
	(0.003)	(0.007)	(0.004)
Suburban	-0.002	0.005	-0.017
	(0.006)	(0.017)	(0.010)
Rural	-0.017*	0.032	-0.018
	(0.007)	(0.018)	(0.012)
Teacher's Age	0.000	0.001*	-0.001
	(0.000)	(0.001)	(0.000)
Teacher has Graduate Degree	-0.004	0.015	-0.005
	(0.005)	(0.013)	(0.009)
Teacher's Experience	0.000	-0.002	0.001*
	(0.000)	(0.001)	(0.001)
Teacher is African American	-0.004	0.013	-0.001
	(0.007)	(0.016)	(0.011)
Teacher is Hispanic	0.013	-0.037	0.028**
	(0.009)	(0.023)	(0.010)
School Start Date (Most Recent Fall)	-0.001**	-0.000	-0.001
	(0.000)	(0.001)	(0.001)
Gender missing	-0.177**	0.000	0.000
	(0.009)	(0.000)	(0.000)
Race missing	0.024	-0.004	-0.010
	(0.014)	(0.038)	(0.022)
Mother education missing	-0.002	-0.011	0.003
	(0.006)	(0.018)	(0.011)
Poverty missing	-0.003	0.001	-0.025
	(0.009)	(0.025)	(0.014)
Teacher age missing	-0.003	0.030	-0.032
	(0.014)	(0.038)	(0.024)
Teacher education missing	0.000	0.068	-0.033
	(0.018)	(0.066)	(0.042)
Teacher experience missing	0.003	-0.082	0.070
	(0.018)	(0.067)	(0.043)
Start date (Most Recent Fall) Missing	-17.169**	-2.200	-14.298
	(3.992)	(11.020)	(7.622)
Constant	17.915**	3.119	15.167*
	(3.991)	(11.023)	(7.623)
N	16,150	4,170	4,020
R-squared	0.037	0.043	0.059

Note: Based on the author's calculations using the ECLS-K without sample weights. Each column represents a separate regression with the number of days in the period indicated as the dependent variable. ** represents estimates statistically significant at the one percent level and * at the five percent level. Standard errors are in parentheses and are clustered at the school level.

Table 3. Effect of a School-Year on the Gains in Test Scores Between Fall and Spring of Kindergarten

	(1)	(2)	(3)
Panel A. Reading Scores			
School Year Period	1.586**	1.441**	1.443**
	(0.013)	(0.200)	(0.233)
Panel B. Math Scores			
School Year Period	1.510**	1.437**	1.408**
	(0.012)	(0.175)	(0.214)
Observations		16,150	
Child Controls?	No	Yes	Yes
School Fixed Effects?	No	Yes	No
Teacher Fixed Effects?	No	No	Yes

Notes: Based on the authors' calculations using the ECLS-K without sample weights. Each column in each panel represents a separate regression with the gains score indicated as the dependent variable. Child characteristics include the child's race, gender, region and area type (urban, rural) of residence and own mother's education. Standard errors are clustered at the school level. ** represents estimates statistically significant at the one percent level and * at the five percent level.

Table 4. Effect of a 250 Day Period at Different Points in Children's Early Elementary Years

	Reading	Math
250 Days: Fall to Spring K	1.586**	1.510**
	(0.013)	(0.012)
Observations	16,150	16,150
250 Days: Spring K to Fall 1st Grade	1.188**	1.126**
	(0.023)	(0.024)
Observations	4,740	4,740
250 Days: Fall to Spring 1st Grade	2.436**	1.716**
	(0.030)	(0.021)
Observations	4,020	4,020

Note: Based on the author's calculations using the ECLS-K without sample weights. Each column and row set represents a different regression with the gains in indicated test score as the dependent variable. ** represents estimates statistically significant at the one percent level and * at the five percent level. Standard errors are clustered at the school level and are reported in parentheses.

Table 5. Effects of School Year Period and Age at Testing for Kindergarteners

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable is Spring Scores in:					
	Reading			Math		
School Year Period	1.543**			1.412**		
	(0.119)			(0.098)		
250 Days of Age		0.341**	0.461		0.514**	0.179
		(0.024)	(0.801)		(0.024)	(0.517)
Observations	13,470					

Note: Based on the author's calculations using the ECLS-K without sample weights. Each column and row set represents a different regression with the gains in indicated test score as the dependent variable. ** represents estimates statistically significant at the one percent level and * at the five percent level. Standard errors are clustered at the school level and are reported in parentheses.

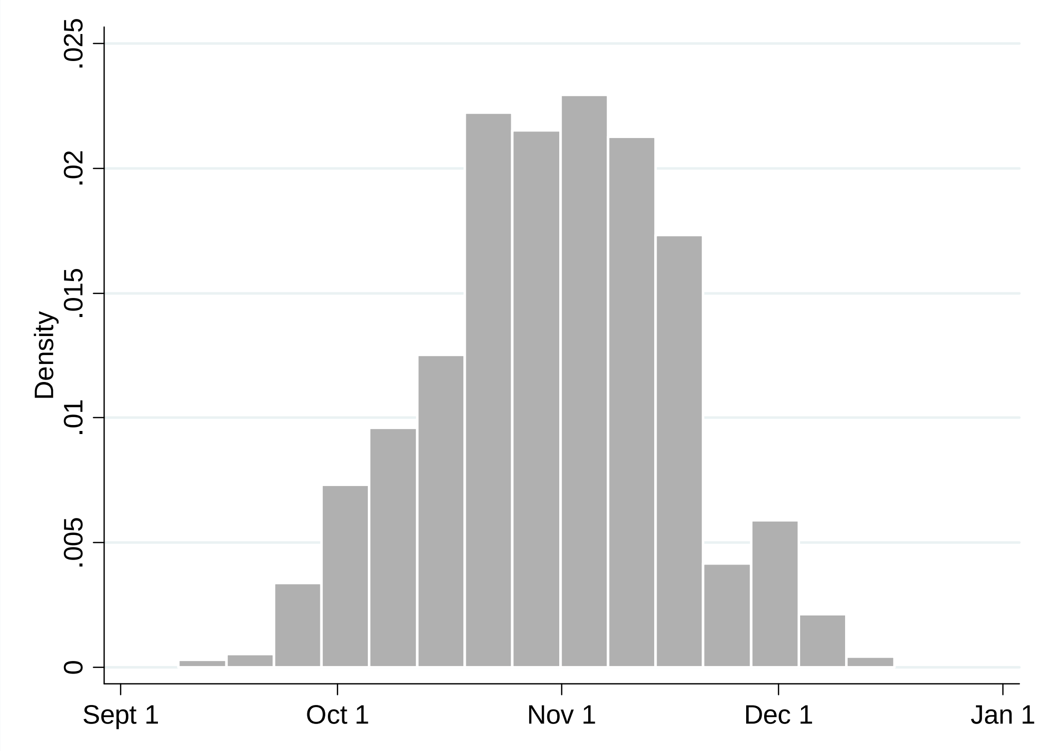
Table 6. Effect of a Half-day versus a Full-day of Kindergarten

	Reading		Math	
	Half-day	Full-day	Half-day	Full-day
Total Days	1.495**	1.660**	1.447**	1.562**
	(0.018)	(0.019)	(0.017)	(0.016)
Observations	7,110	9,040	7,110	9,040

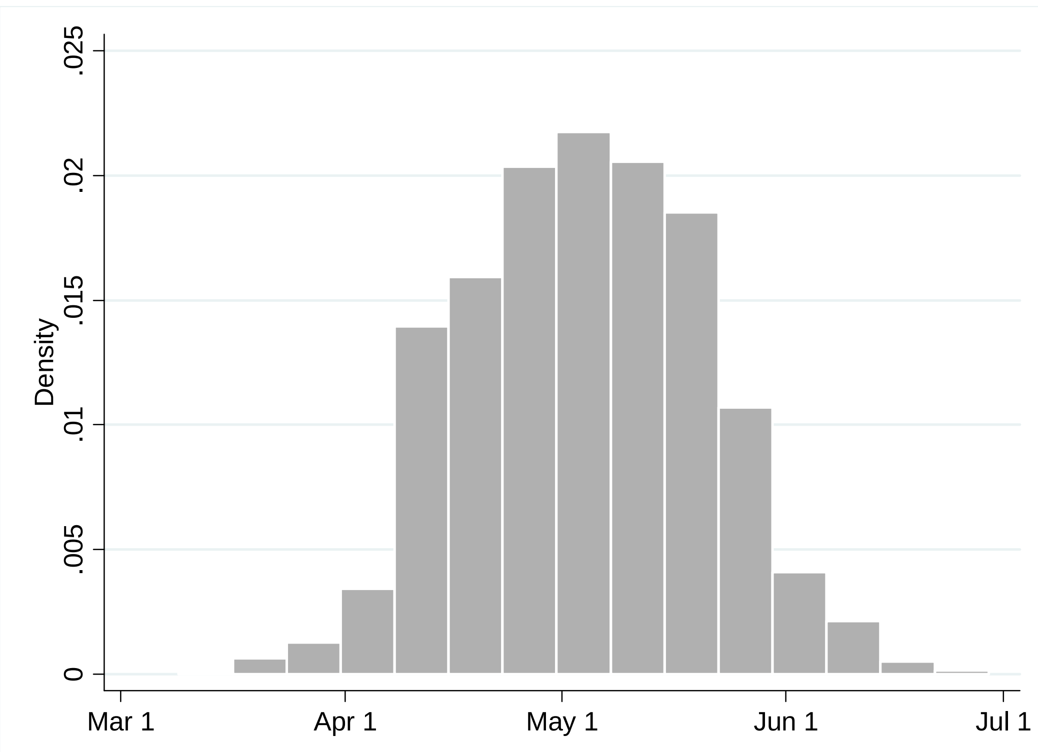
Note: Based on the author's calculations using the ECLS-K without using sample weights. Each column represents a separate regression using the sample of children who enrolled in the specified type of kindergarten. The dependent variable is the gain in the indicated test scores. ** represents estimates statistically significant at the one percent level and * at the five percent level. Standard errors are clustered at the school level and are reported in parentheses.

Figure 1. Distributions of Test Dates

Panel A. Histogram of Fall Kindergarten Test Dates



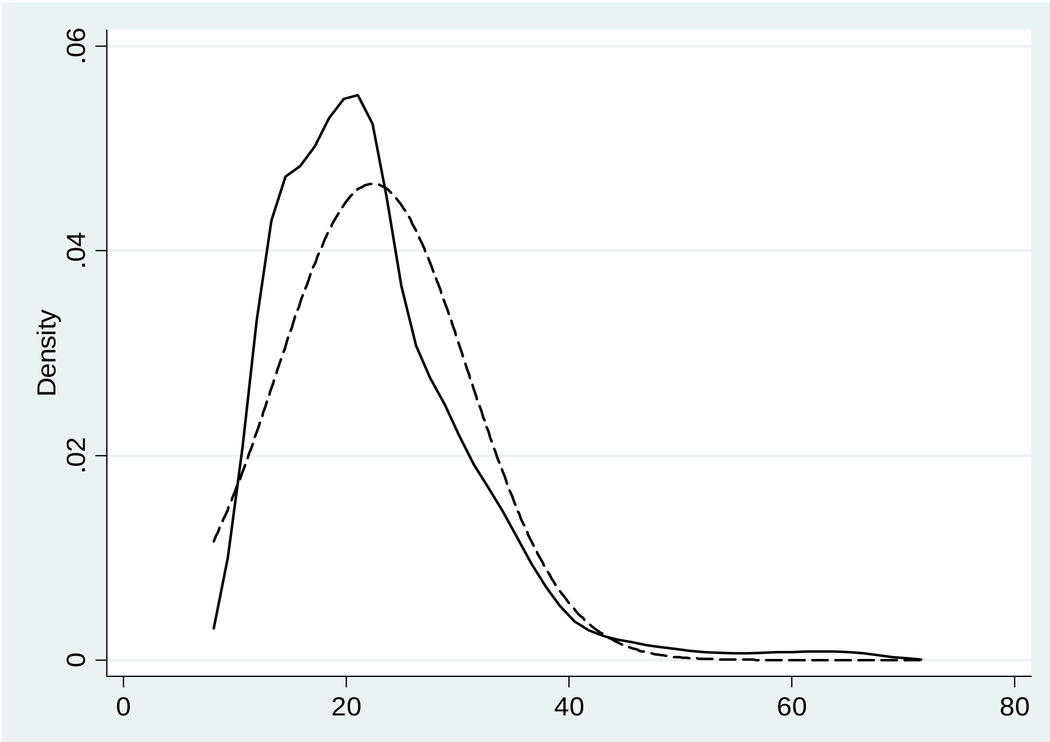
Panel B. Histogram of Spring Kindergarten Test Dates



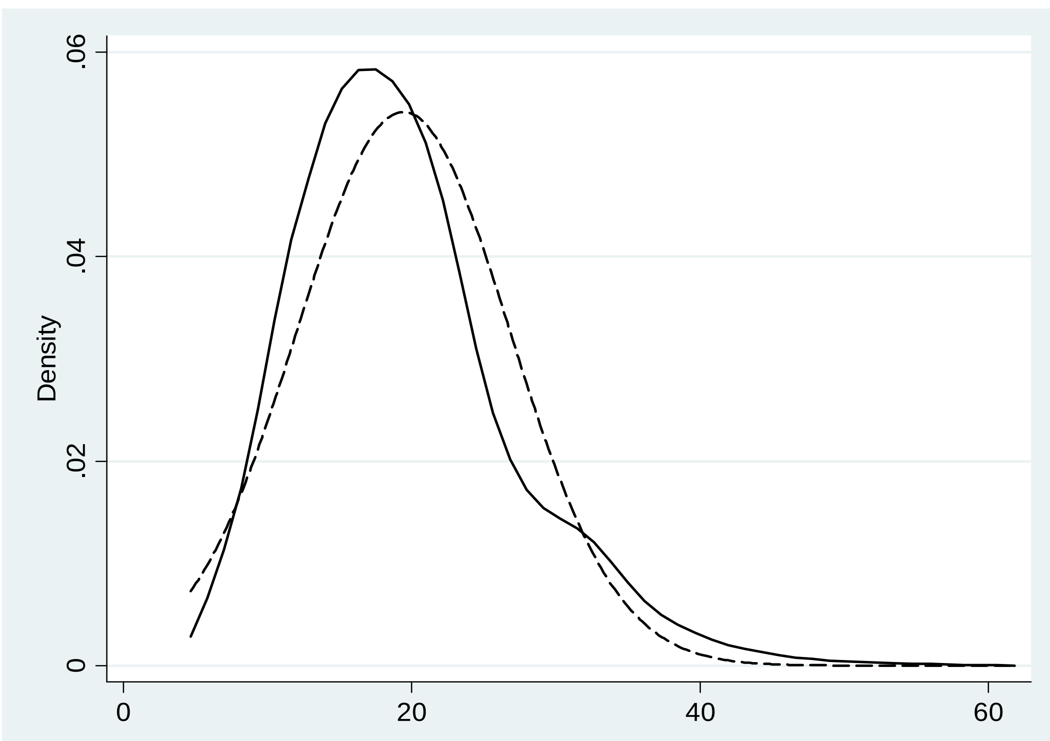
Notes: Based on the authors' calculations using the full unweighted ECLS-K sample described in the text.

Figure 2. Kernel Density Estimates of Kindergarten Test Scores

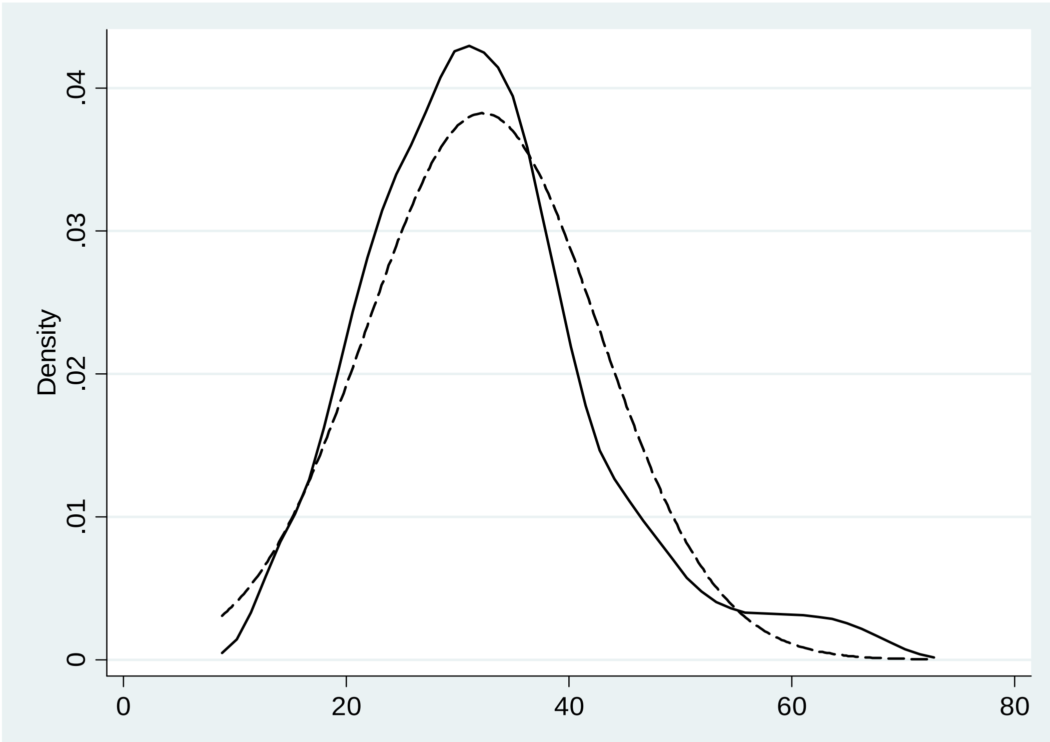
Panel A. Fall Reading Scores



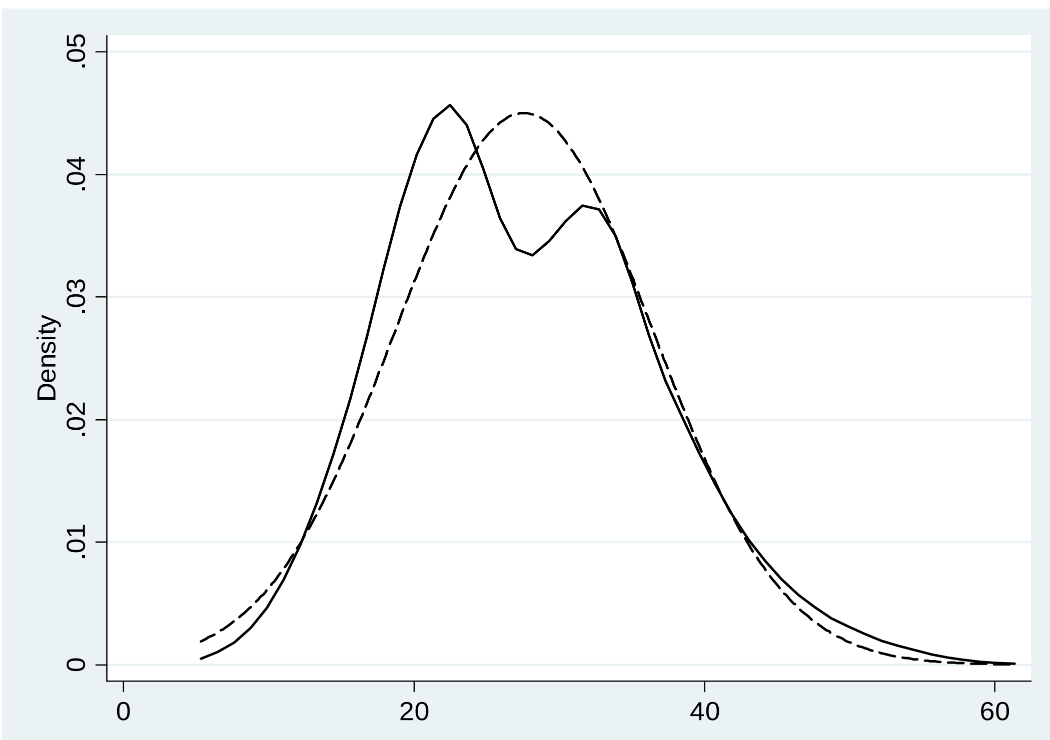
Panel B. Fall Math Scores



Panel C. Spring Reading Scores



Panel D. Spring Math Scores



Notes: Kernels estimated using Epanechnikov kernel and bandwidth of 2.

Appendix A. Language from the ECLS-K Manual

“The field staff was organized into 100 work areas, each with a data collection team consisting of one field supervisor and three assessors. The data collection teams were responsible for all data collection activities in their work areas; they conducted the direct child assessments and the parent interviews, distributed and collected all school and teacher questionnaires, and completed a school facilities checklist.” (page 5-1)

“Once the school administrator agreed to participate, he or she was asked to set an appointment for two visits by the ECLS-K field staff to the school in the fall of the 1998-99 school year. The first visit, the preassessment visit, was to select the sample of children (see section 5.4.2 for more detail on this visit), and the second visit was to conduct the child assessments (see section 5.4.3 for more detail on this visit).” (Page 5-3)

“Beginning in late summer 1998, letters were mailed to school administrators to confirm scheduled visits for the schools. A packet of material was also mailed to the school coordinators asking them to prepare for the preassessment visit to the school. Beginning in September, field supervisors called school coordinators to confirm the dates of the preassessment and assessment visits, answer any questions, and prepare for the preassessment visits. The school coordinators were asked to prepare a list of kindergartners for selecting the sample and to distribute materials such as the study brochure, summary sheets describing the role of teachers and parents in the study, and a letter to teachers to the kindergarten teachers.” (page 5-7)

“For the fall-kindergarten wave, the direct child assessment was administered during a 14-week field period that began in September and ended in early December. In year-round schools, assessment teams made multiple visits to the school to conduct direct child assessments. The assessment team visited the school when each track was in session to assess the sampled children…..When scheduling schools in the fall and the spring, an attempt was made to conduct the direct child assessments at about the same point in time from the beginning of school year and at the end of the year to increase the chances that children’s exposure to instruction was about the same for all children.” (page 5-12)

“In March 1999, letters were mailed to school administrators confirming the scheduled visits for the school that were set up in the fall. Letters were mailed to the school coordinators reminding them of the upcoming visits to the school.” (Page 5-22)

“The direct child assessments were conducted between March and June 1999.” (page 5-26)