



Project
MUSE[®]
Scholarly journals online

RICHARD J. MURNANE
JOHN B. WILLETT
KRISTEN L. BUB
KATHLEEN MCCARTNEY
Harvard Graduate School of Education

Understanding Trends in the Black-White Achievement Gaps during the First Years of School

THE GAPS BETWEEN THE average academic achievement of black and white children have been persistent features of American life. Until quite recently, obvious differences in the school resources provided to children of different races explained substantial portions of these achievement gaps. For example, in 1920 more than one-quarter of the racial gap in children's literacy rates could be explained by differences in easy-to-measure variables such as the school year length and per pupil expenditures.¹

Given the history of blatant discrimination in the school resources provided to American children of different races, it is understandable why the U.S. Congress in the Civil Rights Act of 1964 ordered the commissioner of education to conduct a survey to document "the lack of availability of equal educational opportunities by reason of race, color, religion, or natural origin in public educational institutions at all levels. . . ."² In July 1966 the U.S. Office of Education published the survey results in a 737-page volume entitled *Equality of Educational Opportunity*. Better known as the Coleman Report, named after its lead author, the eminent sociologist James Coleman, this volume documented the substantial gaps between the average mathematics and reading

We thank Roland Fryer for providing the Stata code that he and Steven Levitt used in some of their work on the black-white achievement gap. We also thank Eric Hanushek and Rebecca Maynard for the thoughtful comments on the paper that they provided at the Brookings-Wharton conference.

1. Margo (1986).
2. Coleman and others (1966, p. iii).

achievement of black and white children. However, to the surprise of many educators and civil rights activists, the report found no clear-cut pattern showing that white children attended schools with substantially more of the school resources measured in the survey than did black children. Moreover, school-to-school variation in resources explained very little of the school-to-school variation in children's mathematics and reading achievement. As Harvard government professor Seymour Martin Lipset summarized the results in a conversation with Daniel Patrick Moynihan, "schools make no difference; families make the difference."³

The Coleman Report catalyzed the collection of new data that allowed researchers to challenge the report's findings. Many of the newer data sets provided information on school resources and on children's achievement at more than one point in time. These attributes have allowed researchers to demonstrate conclusively that students learn more in some classrooms and schools than in others. However, with a few exceptions noted below, the newer studies tended to replicate the Coleman Report findings that differences in conventional school resources, such as class size and teachers' educational attainments, do not explain much of the variation in student achievement nor do they explain much of the race-related achievement gaps.⁴

This background provides the context for two provocative papers recently published by Fryer and Levitt.⁵ These economists documented a number of patterns in the relative academic achievement of young black and white children. Their work, which focuses particularly on differences by grade in the black-white test score gap in reading and mathematics, is based on analyses of data on the kindergarten cohort of the Early Childhood Longitudinal Study (ECLS-K), a nationally representative sample of more than 20,000 children who entered kindergarten in approximately 1,000 schools during 1998. Key findings of the two Fryer and Levitt papers include:

—at the beginning of kindergarten, the black-white achievement gap is approximately 0.40 standard deviations in reading and 0.60 standard deviations in mathematics;

—a parsimonious set of family background characteristics explains all of the black-white achievement gap in reading and more than 80 percent of the gap in mathematics;

3. Godfrey Hodgson, "Do Schools Make a Difference?" *Atlantic*, March 1973, p. 35.

4. For a summary of this evidence, see Hanushek (2003).

5. See Fryer and Levitt (2004 and 2005).

—the black-white achievement gap in both reading and mathematics increases by approximately 0.10 standard deviations during each of the first four years of elementary school (kindergarten through third grade);

—there are no important differences between black and white students in the average values of variables that are typically used to measure school quality (for example, average class size and educational attainments of teachers);

—the typical measures of school quality are not statistically significant predictors of students' reading or mathematics achievement. This suggests that reducing class size or requiring that teachers in schools serving large numbers of black students obtain more educational credentials is unlikely to close the black-white achievement gap.

Two aspects of the evidence presented by Fryer and Levitt are especially troubling. The first is that the black-white achievement gaps in both reading and mathematics are much larger at the end of third grade than at the beginning of kindergarten. This suggests (although it by no means proves) that schooling exacerbates inequalities rather than reduces them. The second is that the ECLS-K data appear to shed little light on why the black-white achievement gap becomes larger during the first four years of schooling. The lack of a clear explanation hinders efforts to design policies to alter these patterns.

The research described in this paper has two goals. The first is to examine the extent to which the results reported by Fryer and Levitt are sensitive to model specification. We focused particularly on the question of whether the evidence on the role of school quality is stronger when we fit models that take advantage of the longitudinal nature of the ECLS-K data set. The second goal is to examine whether the patterns that Fryer and Levitt documented in the ECLS-K data set are also present in a smaller but richer longitudinal data set collected with the support of the National Institute of Child Health and Human Development (NICHD). We focused on the question of whether the detailed information present in the NICHD data set on family and school experiences helps us explain the provocative patterns that Fryer and Levitt have documented.

We describe trends in the black-white achievement gap in mathematics and in English Language Arts (ELA). For three reasons we limit the discussion of school resources to their impact on students' mathematics achievement. First, a number of studies have shown that differences in school resources play a larger role in explaining gains in children's mathematics achievement than gains in reading achievement. Second, students' mathematics test scores predict post-graduation labor market earnings better than

do reading scores. Finally, it enables us to keep the number of tables in the paper manageable.⁶

Our paper describes the NICHD data set and documents the steps we took to create an analysis sample from the ECLS-K data set that is comparable to the NICHD data set. We demonstrate that the sample distributions of critical variables common to the two data sets are similar, show that the trends in black-white achievement gaps in the two data sets are different in important respects, and suggest explanations for these differences. We also demonstrate that conventional school resources, defined as those that account for most of school expenditures, do not explain variation in students' mathematics achievement in either data set, and we show that less conventional measures of school quality, including the composition of the student body and how instructional time is spent, are important predictors of student achievement. These last findings suggest promising directions for future research on school-based interventions to close black-white achievement gaps.

NICHD Data Set

One data set we used in this study comes from phases I, II, and III of the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development (NICHD SECCYD). Families were recruited into the NICHD study via hospital visits to mothers shortly after the birth of a child in 1991. These hospitals were located in or near: Little Rock, Arkansas; Irvine, California; Lawrence, Kansas; Boston, Massachusetts; Morganton, North Carolina; Philadelphia and Pittsburgh, Pennsylvania; Charlottesville, Virginia; Seattle, Washington; and Madison, Wisconsin. Of the 8,986 women who gave birth during the sampling period, 5,151 met the eligibility requirements of the study and agreed to be contacted two weeks later.⁷ Using a conditional random sampling method (see below), 2,352 families were subsequently phoned and 1,364 of the families that were called then participated in a home visit one month later for the purposes of data collection.⁸

6. For evidence on the first point see Murnane (1975); on the second point see Murnane, Willett, and Levy (1995).

7. Eligibility requirements include: mother healthy, at least age eighteen, and conversant in English; child healthy, singleton, and not adopted; family not planning to move, residing in a neighborhood that was not extremely unsafe, living within one hour of the research site, and not participating in another study.

8. For a detailed description of the recruitment plan, see NICHD Early Child Care Research Network (2001).

The conditional random sampling plan employed by the NICHD SECCYD was designed to include families from diverse ethnic groups, economic backgrounds, and geographic regions, with varying plans for maternal employment during the child's first year of life. The resulting sample is diverse, with 24 percent of the children being ethnic minority (13 percent African American, 6 percent Latino, and 5 percent Asian, Native American, or other ethnicities), 11 percent of mothers not completing high school, and 14 percent of the mothers being single parents. Mothers in the sample had an average of 14.4 years of education and the average family income-to-needs ratio was 3.6 times the poverty threshold.

NICHD Data Set Strengths

The NICHD data set has several important strengths. First, it contains rich descriptors of the child's family background measured longitudinally at all major assessment points, beginning when the target child was one month old and continuing through third grade. These variables include parental education, parental employment status, family income, household size, and an observational measure of mothers' parenting behavior. Second, children's skills in mathematics and ELA were assessed with a well-established instrument, beginning when the children were age fifty-four months (that is, just prior to kindergarten) and continuing through third grade. Student performance on these tests was reported on a scale constructed using Item Response Theory (IRT). Scores on this scale permit estimation of the determinants of skill growth over time. Third, the data set contains rich longitudinal information on each child's teachers and elementary school classrooms, including measures of the teachers' education and experience, the size and racial composition of the class, and indicators of the allocation of classroom time to different activities (derived from direct observations of the classrooms).⁹

NICHD Data Set Limitations

The NICHD data set also has three important limitations. First, its sample size of 1,364 is considerably smaller than that of the ECLS-K data set. Second, it is not completely representative of the population of all children in the United States. Specifically, the sampling strategy deliberately excluded infants who were twins; mothers who were not healthy, were under age eighteen, or

9. For a detailed description of the NICHD SECCYD, see NICHD Early Child Care Research Network (2001) or (<http://secc.rti.org>).

were not conversant in English; families that planned to move within the next year or give up the child for adoption; families that lived in a neighborhood considered unsafe for visits by the people collecting survey data. A consequence of these exclusion criteria is that the NICHD data set does not include adequate representation of U.S. children born into the most difficult family circumstances. Since black children are more likely than white children to be born into low-income families, estimates of black-white achievement gaps based on the NICHD data set may underestimate the gaps in the population of all U.S. children.

The NICHD data set's third limitation is that it provides no exogenous variation in school resources. This is problematic because resources are not equally distributed nor randomly assigned to students in American school systems. Instead, students with particular characteristics tend to receive different resources than students with other characteristics. For example, students of color tend to be assigned to relatively novice teachers.¹⁰ In many systems, students with learning difficulties are assigned to small classes. A result of this nonrandom assignment of students to resources is that it is very difficult to differentiate the effects that school resources have on student achievement from selection effects.

It is important to note that this third limitation of the NICHD data set also pertains to the ECLS-K data set that Fryer and Levitt used in their analyses. For this reason caution is needed when interpreting relationships between school resources and student achievement in both data sets. However, the availability of longitudinal data on both school resources and student outcomes in the ECLS-K and NICHD data sets provides more opportunities to deal with selection effects than has been the case in prior studies based on less rich data sets. In particular, we can focus on the determinants of growth in student achievement rather than limiting ourselves to explaining the variation in achievement levels at one point in time. Also, the longitudinal data help in identifying selection mechanisms. Indeed, we will show that selection effects explain a counterintuitive pattern in the relationship between class size and student achievement in the ECLS-K data set.

The exclusion from the sample of Latino children whose mothers were not conversant in English makes the NICHD data set quite inappropriate for estimating trends in the achievement of Latino children. For that reason, we

10. See Clotfelter, Ladd, and Vigdor (2005).

do not discuss trends in Latino-white achievement gaps. However, we did retain Latino children in our analytic sample to make our analyses as comparable as possible to those of Fryer and Levitt.

Comparing the ECLS-K and NICHD Data Sets

In an effort to create comparable analytic samples of children from the ECLS-K and NICHD data sets, we applied to the ECLS-K data set five of the eight exclusion criteria that had been applied during the recruitment phase of the NICHD study.¹¹ Specifically, we excluded from the ECLS-K data set those families in which the mother was younger than age eighteen at the birth of the target child or was not conversant in English, those infants who were twins, those not healthy at birth, and those who were adopted. We also excluded from both the ECLS-K and NICHD analysis samples children whose race-ethnicity designation was something other than white, black, or Latino.¹² The resulting sample, which we refer to as the ECLS-K restricted sample, was approximately 14.5 percent smaller than the sample employed by Fryer and Levitt.

We then refitted, in the newly restricted ECLS-K data set, the regression models that Fryer and Levitt report fitting in the full ECLS-K data set. Our results were almost identical to those reported by Fryer and Levitt.¹³ Next, we explored whether the Fryer and Levitt findings were sensitive to the decision to use ECLS-K sampling weights in fitting the regression models. We found that they were not. Finally, we created a series of balancing weights that we applied to individuals in the NICHD data set to match the sample demographically to the ECLS-K restricted (weighted) sample. First, we identified three strata—type of schooling (public versus private), region of the country (Northeast and Midwest versus South and West), and race-ethnicity (white, black, and other)—and obtained sample sizes within cells created by these strata in both data sets. We then created a ratio of children in the NICHD data set to

11. The ECLS-K data set lacked the data to apply three exclusion criteria: family neighborhood was deemed unsafe for research assistants to visit, family lived too far from study base, and family was planning on moving within a year.

12. We excluded from the analysis samples sixty-six children in the NICHD data set and 1,767 students in the ECLS-K data set whose racial-ethnic classification was something other than white, Latino, or black.

13. Copies of these analyses are available from the authors.

children in the ECLS-K data set within each cell. These estimates were used as balancing weights in the computation of all descriptive statistics.¹⁴

Table 1 presents descriptive statistics for the key outcome, family background, and school quality indicators in the restricted ECLS-K analytic sample and in our NICHD analytic sample. Overall, these statistics indicate that the ECLS-K and NICHD data sets are fairly well matched. The percentages of black children in the two samples are similar (14 percent in ECLS-K versus 13 percent in NICHD). However, the percentages of Latino children differ (21 percent ECLS-K versus 7 percent in NICHD). Univariate statistics describing selected family characteristics such as child gender and age at the start of kindergarten are also similar in the two samples. There are some minor differences in the distributions of some of the other family covariates. Specifically, the average birth weight of children in the ECLS-K data set is lower than that in the NICHD data set and the percentage of mothers who were teens at the birth of their first child as well as the percentage who reported being on public assistance are higher in the ECLS-K sample. These differences are not surprising given the criteria used in drawing the NICHD sample.

Measures

In the NICHD data set, time-invariant demographic information was collected via maternal report during home visits when the children were one month old. In addition, time-varying demographic information and information on children's mathematical and English Language Arts (ELA) skills were collected from mothers and children when the children were six, fifteen, twenty-four, thirty-six, and fifty-four months old, and again in the spring of the years when the children were attending first and third grades.¹⁵ Similar information was collected from participants in the ECLS-K data set with one important difference: data collection for the ECLS-K cohort began at the beginning of kindergarten. Table 2 provides descriptive statistics by racial-ethnic group for

14. There are two reasons we examined how sensitive the results of fitting the Fryer and Levitt models with the restricted ECLS-K data set were to the decision of whether to use sampling weights. First, we are not confident about the effectiveness of our procedure for creating balancing weights to make the NICHD sample representative of the weighted restricted ECLS-K sample. Second, the software available to us for fitting models that take advantage of the longitudinal nature of the ECLS-K and NICHD data sets (STATA's XTREG procedures) does not permit estimation with sample weights.

15. For a detailed description of the NICHD SECCYD, see NICHD Early Child Care Research Network (2001) or (<http://secc.rti.org>).

Table 1. Means and Standard Deviations for Outcomes, Family Controls, and School Characteristics in the Restricted ECLS-K and NICHD Data Sets, Kindergarten through Third Grade

<i>Indicator</i>	<i>Descriptive statistics</i>	
	<i>ECLS-K</i>	<i>NICHD</i>
	<i>Mean (SD)</i>	<i>Mean (SD)</i>
<i>Outcome variables</i>		
Math—kindergarten/54 months	21.85 (8.90)	423.41 (20.55)
Math—first grade	55.31 (15.93)	469.96 (15.38)
Math—third grade	85.65 (17.68)	496.59 (13.36)
ELA—kindergarten/54 months	27.70 (9.78)	458.01 (13.93)
ELA—first grade	68.49 (20.26)	482.21 (11.92)
ELA—third grade	108.74 (19.82)	495.05 (11.38)
<i>Race-ethnicity and gender indicators</i>		
Black	.14 (.35)	.13 (.34)
Latino	.21 (.40)	.07 (.24)
White	.65 (.48)	.80 (.40)
Female	.485 (.500)	.479 (.500)
<i>Personal characteristics and family background variables</i>		
Age at kindergarten (months)	67.10 (4.38)	64.32 (3.44)
SES—kindergarten/54 months	.024 (.779)	-.083 (.685)
SES—first grade	.024 (.779)	-.121 (.742)
SES—third grade	.026 (.791)	-.028 (.692)
Children’s books ^a —kindergarten/54 months	75.28 (60.13)	.754 (.431)
Children’s books ^a —first grade	106.94 (143.78)	.754 (.431)
Children’s books ^a —third grade	86.66 (175.24)	.964 (.186)
Birth weight (oz.)	104.84 (43.30)	123.34 (18.37)
Teen mother at first birth	.215 (.411)	.073 (.260)
Mother age 30+ at first birth	.159 (.366)	.393 (.489)
Assistance ^b	.429 (.495)	.097 (.296)
Early maternal sensitivity	n.a.	3.12 (.425)

(continued)

Table 1. Means and Standard Deviations for Outcomes, Family Controls, and School Characteristics in the Restricted ECLS-K and NICHD Data Sets, Kindergarten through Third Grade (Continued)

<i>Indicator</i>	<i>Descriptive statistics</i>	
	<i>ECLS-K</i>	<i>NICHD</i>
	<i>Mean (SD)</i>	<i>Mean (SD)</i>
<i>Class size</i>		
First grade	20.76 (5.13)	21.26 (4.39)
Third grade	21.14 (3.95)	21.44 (4.23)
<i>Master's degree</i>		
First grade	.524 (.500)	.415 (.493)
Third grade	.423 (.494)	.431 (.495)
<i>First two years of teaching</i>		
First grade	.074 (.262)	.056 (.230)
Third grade	.055 (.228)	.086 (.281)
<i>25 percent or more students are black</i>		
First grade	.239 (.335)	.119 (.324)
Third grade	.102 (.303)	.114 (.318)
<i>25 percent or more students are Latino^c</i>		
First grade	.128 (.334)	.021 (.144)
Third grade	.115 (.319)	.032 (.177)
<i>Percent students eligible for free lunch</i>		
First grade	27.09 (27.64)	n.a.
Third grade	30.07 (27.87)	n.a.
<i>Proportion of hour spent on math instruction</i>		
First grade	n.a.	8.31 (10.61)
Third grade	n.a.	9.79 (4.39)

Source: Authors' calculations.

n.a. Not available.

SES = Socioeconomic status.

a. In the ECLS-K data set, the variable *children's books* was measured as the total number of children's books in the home; in the NICHD data set, *children's books* is a dichotomous variable indicating that there were ten or more books in the home.

b. In the ECLS-K data set, *assistance* was a dummy variable indicating whether the mother, child, or both received Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) benefits. In the NICHD data set, *assistance* was a dummy variable indicating whether the family received public assistance, including Aid to Families with Dependent Children (AFDC). In the ECLS-K data set, student composition variables were collected at the school level; in the NICHD data set, student composition variables were collected at the classroom level.

c. In the ECLS-K data set, student composition variables were collected at the school level; in the NICHD data set, student composition variables were collected at the classroom level.

Table 2. Means and Standard Deviations by Race-Ethnic Group for Outcomes, Family Controls, and School Characteristics in the ECLS-K and NICHD Data Sets, Kindergarten through Third Grade

<i>Indicator</i>	<i>ECLS-K</i>		<i>NICHD</i>	
	<i>Black</i>	<i>White</i>	<i>Black</i>	<i>White</i>
<i>Mathematics scores</i>				
Kindergarten/54 months	18.19 (6.38)	23.88 (9.11)	404.89 (22.09)	427.96 (17.00)
First grade	46.87 (12.70)	58.80 (15.83)	457.46 (13.85)	472.99 (14.92)
Third grade	74.10 (17.11)	89.71 (16.25)	485.58 (16.93)	498.87 (11.82)
<i>ELA scores</i>				
Kindergarten/54 months	25.02 (7.84)	28.84 (10.07)	447.97 (10.50)	460.99 (13.82)
First grade	61.17 (18.63)	71.64 (20.10)	471.78 (11.90)	484.53 (11.54)
Third grade	98.56 (19.24)	113.37 (17.99)	484.02 (11.99)	497.55 (10.66)
<i>Family background variables</i>				
<i>Female</i>	.483 (.500)	.483 (.500)	.500 (.501)	.478 (.50)
<i>Age of entry into kindergarten (mos.)</i>	66.89 (4.44)	67.37 (4.36)	64.79 (4.11)	64.23 (3.30)
<i>Family SES</i>				
Early family SES (time-invariant)	n.a.	n.a.	-.490 (.55)	-.008 (.76)
Kindergarten/54 months	-.322 (.711)	.234 (.734)	-.536 (.52)	.061 (.68)
First grade	-.368 (.663)	.234 (.752)	-.543 (.64)	.033 (.72)
Third grade	-.400 (.708)	.237 (.741)	-.343 (.481)	.094 (.71)
<i>Children's books^a</i>				
Kindergarten/54 months	39.23 (38.14)	93.85 (59.83)	.658 (.476)	.765 (.42)
First grade	49.69 (58.12)	132.12 (162.05)	.658 (.476)	.765 (.42)
Third grade	42.01 (153.66)	108.33 (194.20)	.879 (.328)	.988 (.11)
<i>Birth weight (oz.)</i>	95.37 (44.70)	109.11 (40.58)	114.61 (14.63)	124.18 (17.98)
<i>Teen mother at first birth</i>	.405 (.491)	.150 (.357)	.208 (.41)	.051 (.22)
<i>Mother age 30+ at first birth</i>	.069 (.253)	.201 (.401)	.227 (.42)	.454 (.50)
<i>Assistance^b</i>	.735 (.442)	.301 (.459)	.376 (.49)	.043 (.20)
<i>Early maternal sensitivity</i>	n.a.	n.a.	2.66 (.497)	3.22 (.374)

(continued)

Table 2. Means and Standard Deviations by Race-Ethnic Group for Outcomes, Family Controls, and School Characteristics in the ECLS-K and NICHD Data Sets, Kindergarten through Third Grade (Continued)

Indicator	ECLS-K		NICHD	
	Black	White	Black	White
<i>School quality variables</i>				
<i>Class size</i>				
First grade	21.27 (4.52)	20.41 (5.25)	21.76 (6.26)	21.03 (4.05)
Third grade	20.59 (4.02)	21.19 (3.86)	20.79 (4.59)	21.52 (4.18)
<i>Master's degree</i>				
First grade	.571 (.496)	.510 (.500)	.440 (.498)	.431 (.50)
Third grade	.388 (.488)	.448 (.497)	.407 (.493)	.440 (.497)
<i>First two years of teaching</i>				
First grade	.098 (.298)	.067 (.250)	.043 (.20)	.045 (.21)
Third grade	.065 (.246)	.054 (.226)	.123 (.33)	.077 (.27)
<i>25 percent or more students are black^c</i>				
First grade	.511 (.500)	.068 (.252)	.410 (.493)	.076 (.265)
Third grade	.409 (.492)	.051 (.220)	.399 (.491)	.073 (.260)
<i>25 percent or more students are Latino^c</i>				
First grade	.079 (.270)	.043 (.204)	.017 (.131)	.012 (.107)
Third grade	.069 (.253)	.037 (.189)	.012 (.107)	.023 (.150)
<i>Percent children in school eligible for free lunch</i>				
First grade	50.81 (33.28)	19.64 (20.83)	—	—
Third grade	53.67 (30.49)	20.19 (19.38)	—	—
<i>Math instruction</i>				
First grade	n.a.	n.a.	10.04 (13.43)	8.05 (10.33)
Third grade	n.a.	n.a.	9.66 (4.30)	9.81 (4.29)

Source: Authors' calculations.

n.a. Not available.

SES = Socioeconomic status.

a. In the ECLS-K data set, the variable *children's books* was measured as the total number of children's books in the home; in the NICHD data set, *children's books* is a dichotomous variable indicating that there were ten or more books in the home.

b. In the ECLS-K data set, *assistance* was a dummy variable indicating whether the mother, child, or both received WIC benefits; in the NICHD data set, *assistance* was a dummy variable indicating whether the family received public assistance, including AFDC.

c. In the ECLS-K data set, student composition variables were collected at the school level; in the NICHD data set, student composition variables were collected at the classroom level.

mathematics and ELA scores at the beginning of kindergarten and end of first and third grades, as well as for family background variables and measures of school quality for both the ECLS-K and NICHD data sets.

Outcome Variables

At age fifty-four months, and at the end of first and third grades, children in the NICHD data set were administered subscales of the Revised Woodcock-Johnson Psycho-Educational Battery (WJ-R).¹⁶ The WJ-R is a comprehensive set of individually administered tests designed to measure a broad range of cognitive abilities and achievement. The Tests of Achievement assess mastery of broad curricular areas such as reading, mathematics, written language, and general knowledge.¹⁷ The Tests of Cognitive Ability measure factors such as long-term retrieval, short-term memory, processing speed, auditory and visual processing, comprehension, knowledge, and reasoning.¹⁸ We used scores on the Applied Problems subscale of the Tests of Achievement as a measure of mathematics skills. We created a composite of each child's score on the Memory for Sentences and Picture Vocabulary subscales from the Tests of Cognitive Ability as a measure of their ELA skills. Since we analyze measures of children's mathematics and ELA skills obtained at different ages, we used vertically equated IRT-scaled scores as our measures of achievement.¹⁹

Among children in the NICHD data set, the black-white gap in average mathematics scores at age fifty-four months is about 23 points and the gap at the end of third grade is about 13 points (see table 2). Thus as measured in the age-equatable IRT score metric, the average math skill gap declined by 43 percent between the beginning of kindergarten and end of third grade. However, if we follow Fryer and Levitt's practice of measuring the achievement gap in terms of standard deviations in the score distribution at those particular ages, the size of the black-white mathematics achievement gap falls by only 9 percent, from 1.1 standard deviation at age fifty-four months to 1.0 standard deviation at the end of third grade. The reason, as shown in table 1, is that the standard deviation in the NICHD mathematics score distribution at the end of third grade is approximately one-third smaller than the standard deviation in the math score distribution at the beginning of kindergarten. Thus

16. Woodcock and Johnson (1989).

17. Woodcock and Mather (1989).

18. Woodcock and Mather (1989).

19. Rasch (1960).

the trend in the black-white mathematics achievement gap for members of the NICHD data set is very sensitive to the choice of a measurement metric.

The trend in the black-white ELA achievement gap for children in the NICHD data set is quite different from the trend in the mathematics skills gap. Expressed in IRT-scaled points, the gap is about 13 points at the beginning of kindergarten and at the end of third grade (see table 2). Thus in this metric, the size of the gap remains constant between the beginning of kindergarten and the end of third grade. However, this is not the case when the gap is measured using Fryer and Levitt's measurement convention. One reason is that the standard deviation of the NICHD ELA score distribution at the end of third grade is almost 20 percent smaller than the standard deviation in the ELA distribution at the beginning of kindergarten. Thus in Fryer and Levitt's metric, the black-white ELA achievement gap grows from one standard deviation at the beginning of kindergarten to approximately 1.20 standard deviations at the end of third grade.

We next examined the trend in the black-white mathematics skill gap in the ECLS-K data set. We found that it was also sensitive to the choice of measurement metric, although the pattern was quite different from that in the NICHD data set. Measured in the age-equatable IRT score metric, the black-white gap in mathematics scores increased from approximately 6 points at the beginning of kindergarten to approximately 16 points at the end of third grade (see table 2). Expressed as a percentage of the standard deviation at each age level, the black-white mathematics gap increased from 0.64 standard deviations at the beginning of kindergarten to 0.88 standard deviations at the end of third grade.

The trend in the black-white ELA skills gap is more similar to the trend in the math skills gap for the ECLS-K data set than for the NICHD data set. Expressed in the IRT score metric, the black-white ELA achievement gap grows from approximately 4 points at the beginning of kindergarten to 15 points at the end of third grade. Measured in Fryer and Levitt's metric, the gap increases from approximately 0.40 standard deviations to approximately 0.75 standard deviations.

A key part of the explanation for the difference between the two data sets in the black-white mathematics skill gap trend (and to a lesser extent the trend in the ELA gap) is that the tests are quite different. The tests used in the ECLS-K study focus on skills taught in the relevant grades in school. The Woodcock-Johnson tests used in the NICHD study are broad-based measures of skills.

Family Characteristics

Following procedures for the construction of a composite measure of family socioeconomic status (SES) laid out in the ECLS-K data codebooks, we created a time-varying measure of SES in the NICHD data set based on indicators of total family income as well as occupational prestige scores and education of each child's parents. For mothers who were not partnered, only maternal occupational prestige and education (as well as total family income) were included in the SES composite. We then created a time-invariant measure of the child's baseline SES by averaging the values of composite socioeconomic status at ages six months and fifteen months. We subtracted this baseline value from subsequent values of time-varying SES, and included these latter deviations as a time-varying measure of the child's SES in our later regression models along with the baseline measure itself.

One additional family variable we used in selected analyses was a measure of parenting behaviors, specifically maternal sensitivity. Sensitivity was assessed during a mother-child structured play interaction when children were six and fifteen months of age. Mother-child interactions were videotaped in semistructured fifteen-minute observations. Interaction activities included two tasks that were too difficult for the child to carry out independently and required the parent's instruction and assistance. The interaction was rated using seven-point global rating scales, which were designed to capture the mother's emotional and instrumental support for the child during collaborative interactions between mother and child. Our measure of maternal sensitivity is the simple average of the values at the two time periods of a composite of three subscales: supportive presence, respect for the child's autonomy, and hostility (reflected).²⁰ An attraction of this variable is that it provides a direct measure of the quality of stimulation and support very young children receive from their mothers. A limitation is that the variable may be endogenous in that babies' personalities and skills may elicit behaviors from their mothers. For this reason, we discuss in a separate section our exploration of the role this variable plays in predicting children's subsequent skill levels.

As indicated in table 2, the families of black children in the NICHD sample had fewer resources of a variety of types than did white children in this sample. The families of black children had lower socioeconomic status, were

20. For more information on the maternal sensitivity measure, see NICHD ECCRN (1999).

less likely to have at least ten books in the home, and were more likely to be on public assistance. Also, on average, black mothers scored more than one standard deviation lower on the maternal sensitivity measure than white mothers. The black-white differences in family background characteristics are even more stark for children in the ECLS-K sample than for those in the NICHD sample. The likely explanation is that the designers of the NICHD study excluded children who lived in neighborhoods too dangerous for interviewers to visit and black children are more likely than white children to live in such neighborhoods.

School Quality Indicators

In our analyses, we examined the impact on mathematics skills of three types of school quality indicators, each of which has been the subject of public policies aimed at improving school quality. The first are conventional resources, which are things that schools can alter relatively easily, including class size and whether teachers have a master's degree or more than two years of teaching experience. Given that the salaries of almost all teachers in U.S. public schools are based on salary schedules that reward highest degree attained and years of experience, it is not surprising that these conventional resource variables typically explain a large percentage of the variation in school expenditures per student.

There are relatively few differences in either data set in the average level of conventional resources provided to black and white children. As shown in table 2, average class sizes and the percentages of teachers who have master's degrees are about the same for black children as for white children. Thus to look for inequalities in school resources as an explanation for black-white achievement gaps one must look for differences in resources that are more subtle than class sizes and the educational attainments of teachers.

One exception to the general pattern that average levels of conventional resources are about the same for black children as for white children is the likelihood of being taught by a novice teacher. In both data sets, black children are more likely than white children to be taught by a novice teacher in some grades (see table 2). In the ECLS-K data set, this is the case when children are in first grade as well as third grade. In the NICHD data set, it is the case in third grade. This pattern, documented in many studies,²¹ is disturbing because recent high-quality research has shown that novice teachers are less

21. See, for example, Boyd and others (2005); Clotfelter, Ladd, and Vigdor (2005).

effective than those with at least two years of experience in increasing students' achievement.²² For this reason, we include among our measures of school resources a dichotomous variable that takes on a value of 1 if teachers are in their first two years of teaching (zero, otherwise).

The second set of school quality indicators are measures of the racial and ethnic composition of the student body in each student's school (in the ECLS-K data set) or class (in the NICHD data set). The ECLS-K data set only provides information on whether the percentage of black students and percentage of Latino students fell into particular numerical ranges, one of which was greater than 25. Preliminary analyses showed that the important distinction was whether the percentage of black students and the percentage of Latino students were greater than 25. For that reason we included dichotomous indicators for schools that fell in these categories. To keep the model specifications used in the two data sets as similar as possible, we adopted these same variable definitions for the NICHD data set.²³

One reason racial composition measures are interesting is that reducing the racial segregation of American public schools has been a major public policy thrust over the last fifty years. Although analysts differ on the consequences of desegregation policies, several studies have found that children learn less in schools or classes in which a large percentage of the children are black or Latino.²⁴ A likely explanation is that race and ethnicity serve as indicators of poverty and schools serving high concentrations of children living in poverty find it especially difficult to create effective learning environments. To learn something about whether it is racial and ethnic composition *per se* that matters, or whether these variables serve only as proxies for concentrations of poverty, we fitted models using the ECLS-K data set in which we included more direct measures of the concentration of poverty, namely the percentage of children eligible for a free or reduced price lunch, as well as the racial-ethnic student body composition measures.

Not surprisingly, the racial composition of classmates is quite different for black children than for white children (table 2). In both data sets, at least 40 percent of the black children attended schools in which at least 25 percent

22. See, for example, Clotfelter, Ladd, and Vigdor (2006); Rivkin, Hanushek, and Kain (2005); Kane and Staiger (2005); Murnane and Phillips (1981); Rockoff (2004).

23. The results of preliminary data analyses with the continuous measures of the racial-ethnic composition in the NICHD data set were not substantively different from the results reported in the paper that are based on the dichotomous indicators of the racial-ethnic composition of the students in the class.

24. Hanushek and Rivkin (2006).

of their classmates were black, while only about 7 percent of white students attended schools in which at least 25 percent of the students were black. It was also the case that the percentage of children eligible for a free or reduced price lunch was much higher in schools attended by black children (about 50 percent) than the analogous percentage in schools attended by white children (about 20 percent) in the ECLS-K data set (these data were not available in the NICHD data set). These patterns reflect the segregated nature of housing patterns in the United States.

The third set of school quality indicators included in our analyses measures the amounts of time that teachers spent on mathematics instruction. These data, which are available only for the NICHD data set, come from classroom observations conducted by trained observers using the standardized Classroom Observation System (COS) in first and third grades.²⁵ During a three-hour observation cycle, trained viewers recorded data on child and teacher behaviors, activities, and setting quality. More specifically, during two forty-four-minute observation cycles, a time sampling method was used to record discrete codes describing the activities in one ten-minute period of thirty-second observe and thirty-second record intervals. As such, discrete behaviors were sampled for a total of thirty minutes across each of the two observation cycles, for a total of sixty minutes of observation. We created a summary score for the amount of time spent on mathematics instruction by first summing the number of observed behaviors of mathematics teaching across the thirty segments within an observation cycle and then by summing the total observed behaviors across the two cycles. The ensuing measure can be described as the average number of minutes per each hour of observation that the teacher was engaged in teaching mathematics.

The variable measuring amount of school time teachers spend teaching mathematics is interesting because prior studies have shown that the amount of instructional time matters.²⁶ If instructional time does affect student achievement, policies aimed at increasing the amount of time black children spend learning mathematics could contribute to closing black-white achievement gaps. However, there is a need for caution in interpreting the evidence on the relationship between instructional time and children's achievement in this study because instructional time may be endogenous. That is, teachers may devote more time to teaching mathematics, for example, if they find that

25. For more information on the COS, see NICHD Early Child Care Research Network (2002 and 2005).

26. See, for example, Carroll (1963).

children have trouble mastering critical mathematical skills. The estimate of the impact of instruction on achievement from single equation models may be downwardly biased as a result of this potential endogeneity.

As shown in table 2, primary school teachers of children in the NICHD data set allocate an average of eight to ten minutes of each instructional hour to teaching mathematics. There are no large, consistent differences between teachers of black children and those of white children in the allocation of time to mathematics instruction.

Statistical Analyses

We addressed our research questions by fitting random-effects and fixed-effects regression models separately within each data set. In each case, we began with a set of baseline models in which we investigated the relationship between the child's mathematics achievement and the child's age, race-ethnicity, and gender. A regression model typical of those we fitted is:

$$(1) \quad Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 Black_i + \beta_4 Latino_i \\ + \beta_5 (Black_i \times t_{ij}) + \beta_6 (Latino_i \times t_{ij}) \\ + \beta_7 (Female_i) + \beta_8 (Female_i \times t_{ij}) \\ + \beta_9 (Female_i \times t_{ij}^2) + (\varepsilon_{ij} + u_i).$$

In this hypothesized model, outcome Y_{ij} represents the i th child's IRT-scaled mathematics achievement on the j th occasion of measurement; predictor t_{ij} represents the child's age (measured in years and recentered on age of 4.5 years—that is, $t_{ij} = AGE_{ij} - 4.5$); $Black_i$ and $Latino_i$ are dichotomous time-invariant predictors representing the child's race-ethnicity; $Female_i$ is a dichotomous time-invariant indicator that the child is female; ε_{ij} is the usual child- and occasion-specific residual; and u_i represents the time-invariant random effect of the child.

In each data set, we found that the age-trajectories of children's mathematics achievement were consistently a quadratic function of their age. In addition, two-way interactions between ethnicity and linear age made statistically significant contributions to the prediction of mathematics achievement, as did two-way interactions between gender and both linear and quadratic age. Consequently, these terms are included here in our baseline model and in all subsequent hypothesized models below. In the ECLS-K

data set, two-way interactions between race-ethnicity and quadratic age also made statistically significant contributions to the prediction of mathematics achievement. These interactions were included in the regression models fitted with the ECLS-K data set. It is the statistically significant presence of the non-linear impact of child age and the interactions between child race-ethnicity and age that permits the academic trajectories of children of different ethnicities to converge as they grow older, closing the achievement gaps. For each regression model fitted in both data sets, we conducted supplementary general linear hypothesis tests on the apparent gaps in average achievement between black and white children, at age 4.5 years and the end of third grade. Our fitted models and the results of these supplementary tests are reported below for each data set.

In follow-up analyses discussed later, we systematically refitted all hypothesized regression models in each data set, treating the child-specific random effects as fixed effects. In these additional models, all time-invariant effects—such as child ethnicity in the baseline models—drop from the regression model, but important interactions between ethnicity and child age are retained because they are time-varying. Where comparisons could be made, our principal findings did not differ between the random-effects and fixed-effects specifications of the hypothesized regression models. For that reason we provide only the fitted random effects models.²⁷

After fitting the baseline regression models described above and conducting the supplementary tests, we added measures of SES to the random-effects model presented in equation (1).²⁸ Our purpose was to evaluate whether extant differences in children's achievement trajectories by race-ethnicity were, in fact, simply the effect of racial-ethnic differences in socioeconomic status. In these new hypothesized models, we tested for the presence of two-way interactions between the new measures of SES and the other predictors already present in the model, including child age and race-ethnicity. Where they proved important, these interactions were retained and their estimated effects are reported below. We also added to this model the set of time-invariant family background characteristics that Fryer and Levitt included in their models.

27. The results from fitting the models with fixed effects for students are available from the authors.

28. One difference between the regression models fitted in the two data sets is that the model fitted with NICHD data included measures of both time-invariant baseline and time-varying SES (the latter deviated from the baseline value, as described earlier).

Our next step was to add the measure of maternal sensitivity to the model that included family background variables. As explained above, this maternal sensitivity variable was only available in the NICHD data set. We did this to learn whether early parenting behaviors predicted children's mathematics scores several years later, even after accounting for conventional family background variables.

Finally, in each data set we added (sequentially) one of the three kinds of time-varying school quality indicators (described above) to the hypothesized regression models containing the effects of child age, race-ethnicity, socioeconomic status, and other family background variables. We first entered predictors representing conventional school resources. We then replaced these conventional measures of school resources with the measures of the classroom or school racial composition. Finally, we replaced these with measures of the amount of school time spent on mathematics instruction (in the NICHD data set only). In each case, we explored interactions between these new additions and the existing predictors in the model, conducted the supplementary general linear hypothesis tests (described above), and refitted the models treating the hypothesized random effects as fixed. Our findings follow.

Results

Table 3 provides the estimated regression coefficients from fitting the baseline model and the model that includes family background variables for both data sets. It also lists the results from a model fitted with the NICHD data set that includes the maternal sensitivity variable and its interactions with time. Random effects for children are specified in these models. Since the estimated coefficients on these control variables hardly change when the school resources variables are added to the model, we do not report these coefficients in subsequent models. The first panel in table 4 lists the coefficients on the three sets of school variables for the ECLS-K sample. The second panel in the table provides the analogous information for the NICHD sample.

Table 5 summarizes, for both the ECLS-K and NICHD data sets, the predicted sizes of the black-white mathematics and ELA skill gaps from random effects models that do not and that do include family background variables including socioeconomic status. The relevant entries in table 5 (-0.599 and 2.46) show the pattern that Fryer and Levitt emphasized, namely, that in the ECLS-K data set, there is no statistically significant black-white skill gap in mathematics and ELA at the beginning of kindergarten after taking into

Table 3. Results of Regressions Predicting Children's Math Scores Using the Restricted ECLS-K Data Set and the NICHD Data Set, with Random Effects for Individual Children

<i>Variable</i>	<i>ECLS-K</i>			<i>NICHD</i>		
	<i>Baseline</i>	<i>All family background controls</i>	<i>Baseline</i>	<i>All family background controls</i>	<i>Baseline</i>	<i>Early maternal sensitivity</i>
Age	26.20****	26.34****	24.84****	25.12****	36.79****	
Age ²	-2.06****	-2.09****	-1.80****	-1.85****	-3.49****	
Age at entry to kindergarten	.444****	.487****	-5.55****	-6.38****	-6.66****	
Black	-5.57****	-.599	-20.31****	-12.77****	-8.65****	
Latino	-6.23****	-1.83****	-11.39****	-7.95****	-6.26****	
Black × age	-4.46****	-4.30****	1.91****	1.07****	.554	
Latino × age	-2.22****	-2.08****	1.53****	1.10****	.916**	
Black × age ²	.458****	.434****	—	—	—	
Latino × age ²	.280****	.246****	—	—	—	
Female	.118	.108	4.64****	4.38****	3.60****	
Female × age	-1.03****	-1.09****	-4.27****	-4.16****	-3.93****	
Female × age ²	.060*	.078**	.663****	.655****	.615****	

Early SES (time-invariant)	n.a.	5.16****	3.47****
Early SES × age	n.a.	-1.18****	-791****
SES (time-varying)	3.46****	2.51****	2.07****
Children's books	.003****	7.12**	4.60
Birth weight (oz.)	.017****	.043**	.034
Teen mother at first birth	-2.18****	-3.26*	-2.12
Mother age 30+ at first birth	1.92****	2.04**	1.36*
Assistance	-3.65****	-5.14****	-2.83**
Early maternal sensitivity			12.88****
Early maternal sensitivity × age			-3.70****
Early maternal sensitivity × age ²			.524****
<i>Variance components</i>			
σ^2_u	111.94	139.95	110.46
σ^2_e	72.93	97.61	93.70

Source: Authors' calculations.

n.a. Not available.

SES = Socioeconomic status.

**** $p < 0.001$, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 4. Results of Regressions Predicting Children's Math Scores Using the Restricted ECLS-K Data Set and the NICHD Data Set, with Random Effects for Individual Children

	<i>Conventional school resources</i>	<i>Racial-ethnic composition of student body</i>	<i>Racial-ethnic composition and percent students in poverty</i>
<i>ECLS-K data set</i>			
Control variables ^a	✓	✓	✓
Class size	.065***		
Master's degree	-.025		
First two years of teaching	-.350		
25 percent or more students are black		-.741**	-.011
25 percent or more students are Latino		-.705**	.004
Percent students eligible free lunch			-.030****
<i>Variance components</i>			
σ^2_U	85.56	88.74	88.55
σ^2_e	69.72	71.06	70.73
	<i>Conventional school resources</i>	<i>Racial-ethnic composition of student body</i>	<i>Math instruction</i>
<i>NICHD data set</i>			
Control variables ^a	✓	✓	✓
Class size	-.004		
Master's degree	-.256		
First two years of teaching	-.074		
25 percent or more students are black		-1.22	
25 percent or more students are Latino		-2.65	
Math instruction			.053*
<i>Variance components</i>			
σ^2_U	114.70	114.70	115.99
σ^2_e	95.65	95.84	95.65

Source: Authors' calculations.

**** $p < 0.001$, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

a. The control variables include all of the child and family covariates listed in columns 2 and 4 of table 3. In addition, we controlled for whether or not the child was held back in kindergarten, first, and third grade, as well as the interaction between held back in kindergarten and age.

account the effects of differences in observed family background characteristics. The relevant entries in table 5 (-12.77 and -7.68) show that the patterns are quite different in the NICHD data set. Even after taking into account the effects of observed family background characteristics, there are substantial black-white skill gaps in mathematics and ELA at the beginning of kindergarten for children in the NICHD data set. The likely explanation is that the Woodcock-Johnson tests measure a broader range of skills than do the tests administered to the ECLS-K sample.

As shown in table 5, the trend in the fitted black-white mathematics and ELA skill gaps for children in the ECLS-K data set are as reported by Fryer

Table 5. Predicted Black-White Gap in Math and ELA Scores at the Beginning of Kindergarten and End of Third Grade in the Restricted ECLS-K and NICHD Data Sets

Score	ECLS-K		NICHD	
	Beginning of kindergarten	End of third grade	Beginning of kindergarten	End of third grade
Mathematics				
No controls	-5.58****	-15.72****	-20.31****	-12.00****
Family covariates	-.599	-10.47****	-12.77****	-8.12****
Family covariates plus maternal sensitivity	n.a.	n.a.	-8.65****	-6.24****
ELA				
No controls	-3.77****	-15.58****	-12.75****	-12.97****
Family covariates	2.46****	-13.38****	-7.68****	-9.63****
Family covariates plus maternal sensitivity	n.a.	n.a.	-5.17****	-6.93****

Source: Authors' calculations.

n.a. Not available.

**** $p < 0.001$, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

and Levitt: they grow markedly between the beginning of kindergarten and end of third grade. This is true both in models that control for family background and those that do not. In contrast, the trend in the fitted black-white mathematics skill gap for the NICHD data set, as measured by IRT-scaled scores on Woodcock-Johnson tests, declines by 40 percent between the beginning of kindergarten and end of third grade and the ELA skill gap remains stable in size (see table 5). Again, the likely explanation for the striking difference in trends is that the Woodcock-Johnson tests measure somewhat different skills than do the tests administered to the ECLS-K sample. It is also important to keep in mind that trends in the size of the black-white achievement gap depend on the measurement metric. We report here trends in test score points on the vertically equatable IRT test scale metric. The patterns are somewhat different when the gaps are measured in terms of standard deviation of the test score distribution measured at that particular age—the metric Fryer and Levitt used.

Maternal Sensitivity

As explained above, the maternal sensitivity variable that we constructed for the NICHD data set provides a direct measure of the quality of stimulation and support very young children receive from their mothers. The variable has the potential to shed light on the importance of children's early at-home experiences in determining later school success.

We recognize that there may be problems of endogeneity with the measure of mother's parenting behavior. That is, we have no way of knowing whether sensitive mothers produce higher-performing children or whether higher-performing children evoke sensitive responses from their mothers. By including sensitivity measures from a very early age (an average of the scores taken when children were six and fifteen months old), we attempted to reduce the endogeneity, though we cannot be sure that we have eliminated it.

To explore whether the maternal sensitivity variable predicts subsequent math skills for children in the NICHD data set, we added the measure of maternal sensitivity and its linear and quadratic interactions with time to the model that included conventional family background variables. The results are reported in the far right column of table 3.

Children whose mothers scored highly on the measure of maternal sensitivity performed much better on subsequent tests of mathematical skills than did children whose mothers had lower scores on this measure. Since black mothers, on average, scored more than one standard deviation lower on the maternal sensitivity measure than did white mothers, it follows that controlling for this variable reduces the predicted size of the black-white gap in mathematics skills. As indicated in table 5, net of the impact of conventional family background characteristics, the predicted black-white mathematics score gap at the beginning of kindergarten is 12.77 IRT-scale points. In the model that also controls for maternal sensitivity, the predicted black-white mathematics skill gap is 8.65 points. In other words, differences in early parenting behaviors explain approximately one-third of the black-white gap in mathematics skills at the beginning of kindergarten that remains after taking conventional family background characteristics into account. The impact of early maternal sensitivity on mathematics skills declines somewhat with age. However, as shown in table 5, differences in early parenting account for almost one-quarter of the black-white mathematics skills gap at the end of third grade that remains after taking conventional family background variables into account.

Differences in early parenting behaviors are equally important in explaining that part of the black-white gap in ELA skills that remains after taking conventional family background characteristics into account. As shown in table 5, differences in early parenting behaviors explain approximately one-third of the remaining ELA gap at the beginning of kindergarten and slightly more than one-quarter of the remaining ELA gap at the end of third grade.

The results on the role of early parenting behavior in predicting children's subsequent mathematics and ELA skills have two implications. First, they

support the argument of child psychologists that children's experiences during their first years of life have lasting impacts. Second, the results show that measures of children's mathematics and ELA skills measured at the beginning of kindergarten and at the end of first grade do not capture the full influence of early childhood experiences. As Todd and Wolpin (2003) explained, this pattern illustrates the difficulty of developing models that accurately estimate the impact of home and school influences on children's achievement.

Conventional School Resources

As shown in table 4, there was no evidence from either data set that teachers with a master's degree are more effective in enhancing students' mathematics skills than are teachers without this degree. This pattern is consistent with the results of the vast majority of prior studies. It also makes sense because the requirement imposed by many states that teachers acquire a master's degree within their first few years of teaching has created incentives for the creation of a large number of relatively undemanding, low-quality master's degree programs.

We also do not find any statistically significant evidence in either data set that novice teachers are less effective than those with at least two years of experience (see table 4). Differences in the design of both the ECLS-K and the NICHD data sets are the likely explanation for why our results differ from those of several recent studies that do show that individual teachers become more effective as they gain experience. Neither the ECLS-K nor the NICHD studies follow teachers over time. For that reason our estimates compare the performances at one point in time of teachers who have different levels of experience. Selective attrition of the most-effective teachers from the classroom would lead us to underestimate the extent to which teachers improve their performance over their first several years in the classroom.²⁹

A surprising pattern is that class size is positively associated with student achievement in the ECLS-K data set (see table 4). We believe that the explanation for this counterintuitive finding is the process by which students are assigned to classes. We found strong statistically significant negative relationships between class size for students in grade t and their reading and math scores in year $t-2$. In other words, schools tended to place struggling students in smaller classes. Even with the benefits of the smaller classes, these students made less progress than the more-skilled students in larger classes. This

29. For a discussion of this issue and evidence of the impact of research design on results, see Murnane and Phillips (1981).

illustrates the difficulty of isolating productivity effects of school resources in settings in which there is nonrandom assignment of students to resource levels.

Student Body Composition

Estimates based on the ECLS-K data set show that students in schools in which more than 25 percent of the student body consists of black or Latino students learn less mathematics than students in schools with a lower percentage of students of color. The coefficients in the regression models fitted with the smaller NICHD data set show the same pattern although they are not statistically significantly different from zero.

An interesting question is whether the racial-ethnic composition of the student body affects student learning or whether these easily observed variables stand in for less commonly observed variables measuring the percentage of students living in poverty and coming to school with especially great learning needs. To examine this question, we fitted a regression model in the ECLS-K data set that included the percentage of students eligible for free or reduced-price lunch as well as the two indicators of the racial-ethnic mix of the student body. The results are reported in the far right column of the top panel in table 4. The coefficients on the racial-ethnic mix of the student body are not statistically significantly different from zero in this model, but the percentage of students living in poverty is a strong, statistically significant negative predictor of student achievement. This supports the position that the important characteristic of the student body is the percentage of students living in poverty. Schools serving high concentrations of students from poor families face especially large challenges, ones that relatively few schools have been able to consistently master.

Instructional Time

In the NICHD data set we found that the more classroom time teachers devote to teaching mathematics, the more mathematics student learn during the school year. A causal interpretation of the coefficient implies that an increase of ten minutes of math instruction during each hour of the school day (equivalent to an increase of fifty minutes per day in a five-hour school day) would result in an increase in 0.58 points in each student's math score at the end of the school year.³⁰ Over four years of school (from the beginning

30. As we explain earlier in this paper, the presence of selection effects may have led us to underestimate the impact of instructional time on achievement. The logic is that teachers whose children are having difficulty with mathematics may devote more time to teaching it.

of kindergarten to the end of third grade) the predicted increase would be 2.32 points, approximately 20 percent of the black-white gap in math skills at the end of the third grade. Another way of expressing the magnitude of the effect is that closing the 12 point black-white achievement gap in math achievement at the end of third grade (approximately 0.90 standard deviations) would require that black students receive four to five hours more math instruction per day over the first four years of school than white students receive. We interpret our results as indicating that providing more instruction is a possible policy approach for closing the black-white achievement gap. However, it is also important to invest in improving the quality of instruction so that the payoff to each hour of instruction is larger.

Discussion

The results of our examination of black-white mathematics and ELA achievement gaps in the ECLS-K and NICHD data sets shed light on several aspects of Fryer and Levitt's provocative findings. Perhaps most important, their finding that a relatively small set of background characteristics explains almost all of the gaps in mathematics and ELA skills at the beginning of kindergarten appears to stem from the narrow focus of the tests in the ECLS-K data set. In the NICHD data set, substantial black-white gaps in mathematics and ELA skills are present at the beginning of kindergarten even after accounting for virtually the same set of family background characteristics that Fryer and Levitt used in their studies. Moreover, studies using other data sets also report large black-white skill gaps at the beginning of kindergarten.³¹

We also show that the substantial growth in the black-white mathematics and ELA skill gaps over the period from the beginning of kindergarten that Fryer and Levitt present in the ECLS-K data set are not present in the NICHD data set. Measured in IRT-scaled points, the black-white skill gap in mathematics for children in the NICHD data set declines markedly during the first four years of school. The gap in ELA skills does grow, but at a much slower rate than in the ECLS-K data set. We also show that trends in the gap are sensitive to the choice of measurement metric.

We find that a measure of parenting skills constructed from observational data when the child was six and fifteen months old is a strong predictor of children's mathematics and ELA skills during the first years of school. Differences

31. Rouse, Brooks-Gunn, and McLanahan (2005).

in the observed measure of parenting skills account for more than one-quarter of the black-white skill gap that is present among children with the same conventional family background characteristics.

Turning to school resources, the one difference in the distribution of conventional school resources is that black children are much more likely to be taught by novice teachers than are white children. This pattern, which is present in both data sets, is disturbing because a number of recent studies with strong designs show that teachers are much less effective in their first two years of teaching than they are after acquiring at least two years of experience.

We find no evidence supporting the argument that across-the-board cuts in class size are an effective strategy for improving school quality. In interpreting this evidence, it is important to keep in mind that the class sizes in our data set average twenty-one to twenty-three students. Thus our evidence sheds no light on the role of class size in influencing student achievement in classes with thirty-five or forty-five students.

We also find no evidence that requiring teachers to earn master's degrees is an effective strategy for improving school quality. As explained above, a likely explanation is that the master's degree requirement stimulated the creation of many low-quality master's degree programs.

Our evidence based on the NICHD data set is that the amount of time teachers devote to mathematics instruction influences how much mathematics children learn. This suggests the importance of focusing school policy on strategies to improve the quantity (and quality) of instruction children receive in school. Designing public policies to increase the quantity and quality of instruction children receive at school is more difficult than designing policies to reduce class size or to mandate that all teachers earn additional educational credentials. Evidence from schools that have produced student achievement gains by increasing the quality and quantity of instruction shows that schools can play an important role in increasing the achievement of black students.³² However, going to scale with instructional improvements remains an enormous challenge in the schools most of the nation's black children attend.

A final pattern is that student body composition matters, and the percentage of students living in poverty is a better indicator of the challenges schools face in enhancing student achievement than is the racial-ethnic composition of the student body. Altering the housing patterns that produce schools segregated by race and income is a challenge the United States has never been

32. See Levy and Murnane (2004, chap. 8).

willing to embrace. Closing the black-white achievement gap when schools remain segregated by race and income is extraordinarily difficult. To succeed, schools that serve concentrations of poor children must be staffed with skilled, experienced teachers who have learned to work together to provide large amounts of consistent, coordinated, high-quality instruction. Closing the gap is the greatest educational challenge facing the United States today.

Comments

Eric Hanushek: This is a high-quality paper of the type that I have come to expect from these researchers. More importantly, they have turned their attention to a set of intellectual issues that have quite large policy importance. Murnane and his colleagues have picked up on Fryer and Levitt's very provocative analyses, which could be interpreted as suggesting that schools contribute to a growing racial gap in achievement. Murnane and others' analysis in this paper suggests that Fryer and Levitt's data and analysis are in question.

The most important Fryer and Levitt findings (in simplest terms) are that schools appear to contribute to a widening black-white achievement gap, but obvious school policies show little hope for improving the situation. Given the continued policy concerns about the distribution of a student outcomes, these findings that take the focus to the earliest school experiences lead to a new round of questions about what options are available.

Fryer and Levitt's unique study, however, needs some corroboration before policy is based on it. Specifically, much of the evidence is indirect—that differences in school achievement can be explained by neither measured differences between family background characteristics nor school factors, even though preschool differences can be explained by these factors. Further, the results could depend on the measures and samples of a new effort to study achievement—the Early Childhood Longitudinal Study (ECLS-K). This new education sample may in itself have some peculiarities.

This background motivates the study by Murnane and others. They have a simple but important research plan. First, they introduce a new and highly detailed data set from the National Institute of Child Health and Human Development (NICHD) in order to study the sensitivity of the results to the ECLS-K sample. Second, they investigate the sensitivity of the results to various aspects of the model specification.

This activity is extraordinarily valuable. There are many examples of alleged facts that come from one study or one sample proving to be much

less than facts when they are overturned by another study. This situation is especially problematic when policies are swirling around as they are in the area of the racial achievement gap and there is the possibility of making premature policy judgments. The authors of this paper show the merits of careful and imaginative replication. Their findings put a different twist on the racial gap. First, while they confirm the finding that black students enter school less prepared than whites, they find that the early achievement gap looks larger and less explicable than that portrayed by Fryer and Levitt. But second, they find that the gap shrinks with time in school, as opposed to expanding. Third, in the category of old results, they confirm that measured attributes of teachers and schools have little to do with either achievement or the racial gap (a Fryer and Levitt finding). Black and white students face roughly the same measured characteristics of schools, but they are not important in explaining achievement.

Murnane and his colleagues at the same time fail to confirm two common findings in previous work. They do not find that having a rookie teacher is a particular detriment, even though black students tend to get new teachers a disproportionate amount of the time. Further, these authors do not see that the more-segregated schools that black students are likely to attend are a particular disadvantage *per se* (although their measure of racial composition is quite imprecise). Instead, they find that racial composition is simply an imperfect proxy for the aggregate composition of student family income—that is, it is student socioeconomic status and not race.

Finally, the authors do find support for the idea that time on task has a positive effect. To people outside of education research, this seems like something that does not need to be explained. Nonetheless, this seemingly trivial finding has been difficult to document in a convincing way, and even here has some uncertainty surrounding it because Murnane and his colleagues have trouble sorting out the underlying causal mechanism.

Their work leaves a number of puzzles and suggestive ideas. They show quite convincingly that the differences in the samples do not appear to be the cause of the differing results. They do a very careful job of matching samples, and the overall conflict of results remains. So what lies behind the differences? They are led to concentrating on possible differences in test measurement, although it seems like more analysis is needed in this area. While this is plausible, there seem to be many questions on that, and there is no direct evidence.

They do point out clearly that there are alternative test metrics that can be used and that the choice can affect the results. Specifically, it is possible, and common, to translate everything into standard deviation units of performance

(effect sizes), but one can also use the Item Response Theory scaling that puts different levels on the same scale. In the authors' work, the choice can change the conclusions in noticeable ways. Yet the choice is rather arbitrary without any external valuation of educational outcomes. Indeed, while the authors here do not linger on this, the topic of measurement scales for student performance is one that will undoubtedly receive increased attention over time.

The analyses of specific resources and factors affecting performance, while confirming much of the past work on specific resource effects, would seem to need more work. The authors have a small sample that lacks much detail on schools and teachers. In addition, they have trouble telling a compelling story that they have identified causal influences. Because of these limitations, this portion of their study presents the most uncertainty (and probably cannot be resolved with their data).

Perhaps the biggest question revolves around the differences between models of math gaps and English Language Arts (ELA) gaps. Specifically, the authors' ability to explain preschool gaps and the pattern of these gaps in school differs significantly across tests. Essentially, in their new data, the math gap significantly declines with schooling, but the ELA gap is constant or somewhat increasing. While it is possible to tell stories about why these different subject areas might respond differently to parental and school inputs, they all seem to require a lot more analysis.

There are two relatively new and intriguing findings. First, Murnane and his colleagues find powerful and continuing effects of early parenting behavior. This work is starting to dig further into what actually takes place in the family, and it is a welcome line of inquiry. Too much research stops at just finding that family income is correlated with student outcomes and fails to tell us anything about that relationship.

Second, the authors underscore the importance of time on task for determining achievement. This finding hopefully will spur more work in the area. It comes up in many of today's policy debates. For example, recent questions about federal accountability (for instance, No Child Left Behind) have centered on the idea that accountability in a few areas may lead to changes in the emphasis of schools. In particular, schools desiring to improve math and reading performance might spend more time on these subjects. The findings of this paper indicate that this might be an effective strategy by schools. Of course, part of the discussion lingers on whether or not this is a good idea, because spending more time on math implies spending less time on some other things, given an overall time budget. These extended ideas about the

pluses and minuses of more time unfortunately go beyond what Murnane and his colleagues can analyze within their data.

Rebecca Maynard: Murnane and others have undertaken a very careful exploration of the size and character of early differences in math and reading achievement of young children, with a particular focus on differences in the average achievement levels of blacks and whites. This research is important for several reasons. First, as a group, black children (as well as children from most other minority race or ethnic groups) consistently underperform academically relative to white children. Second, the differences in math and reading achievement between black and white children are sufficiently large to matter in terms of the economic and social prospects children face when they reach adulthood. The empirical evidence is quite clear that raising the academic performance of black children is critical for reducing disparities in earnings. Third, whatever worked to reduce achievement differences in the years following the release of the Coleman Report no longer worked by the 1980s.

This paper is a small but very important contribution to understanding the causes of the residual differences in achievement and for identifying promising policy or practice changes that could increase the academic success of black children without compromising that of white children. Three highlights from the paper by Murnane and others follow.

First, this study underscores the importance of identifying appropriate measures of the skills that are important for children of various ages. There is no common currency for measuring academic achievement and, indeed, the currency used affects quite substantially the conclusions that follow. The achievement test itself matters in so far as different tests measure different skills. Yet there is no obvious basis for preferring one test to another. For example, before one can make real meaning out of the fact that the results for reading and math are more similar for the Early Childhood Longitudinal Study (ECLS-K) study sample than for the National Institute of Child Health and Human Development (NICHD) study sample, one should understand how well the achievement tests used in each study measure outcomes that are important for subsequent learning and lifelong success.

It also is important to better understand the interpretation of the metrics for judging achievement. Indeed, as pointed out by Murnane and his colleagues, standardized means provide a quite different picture of changes in black-white achievement gaps over the early grades than do scale results based on Item Response Theory. Standardized mean differences, for example, tell something about where the mean performance of one group of children stands

relative to the overall distribution of performance levels. They tell nothing about performance differences relative to any particular standard of achievement. Moreover, standardized mean differences are highly sensitive to changes in the distribution of scores over time. Raw scores pose different issues. For instance, while they order children in terms of their overall performance, they do not incorporate judgments about whether equal size gains are more valuable at one end of the distribution than another.

A second contribution of this study relates to its findings on the value of increased instructional time. It seems clear from the study results that increased instructional time alone could make an important, but quite modest, contribution to closing the achievement gap. Murnane and others project that four or five hours of math instruction a day over the first four years of school would be required to eliminate the black-white achievement gap in mathematics. This, of course, assumes that the marginal relationship between math instructional time and achievement based on the current range of practice could be extended to a quite different profile of the use of class time, where almost the full school day would be devoted to math instruction or the number of hours in the school day would be extended substantially. Possibly, if increased instructional time were paired with more effective curricula and with better use of the before- and after-school hours, large gains in performance could be achieved. Murnane and others were not able to assess the benefits of such strategies with their data sets. However, there is an emerging literature—which is being enhanced through a number of ongoing randomized controlled trials initiated by the Institute for Education Sciences, U.S. Department of Education—that will provide evidence on the potential benefits of such strategies.

The study's third contribution relates to the clear reminder that schools play a very important role in promoting social justice and economic opportunity, but schools are not the sole solution to social and economic inequities. Much of what matters for the health and well-being of children occurs outside of the school setting—in the home or neighborhood. Murnane and others found evidence that parenting matters. This is important because there is evidence that policy interventions (for example, some home-visiting programs) can improve parenting. Furthermore, it might be possible to improve parenting through strategies involving media campaigns, community education programs, or parenting education classes for high-school students.

Murnane and others also note, but do not pursue, the fact that poverty, not race, seems to be the more important determinant of academic achievement. Therefore both housing and employment as well as training policies are poten-

tially more important than education policies for reducing the achievement gap. Historically, however, policies in these areas have not been designed with any attention to their implications for educational outcomes for children.

This paper suggests three areas for further research. One area relates to understanding better what leads to low achievement, possibly through closer examination of so-called *defiant* children. For example, what is special about low-income, black children who succeed and what is special about those non-poor, white children who fail?

A second area for further study relates to who is having children and how a parent influences a child's achievement. More than 400,000 children are born to teenagers annually and one-third of all children (including a large majority of those born to teens) are born to single mothers. Both teen and single parenting affect the family and neighborhood contexts in which children are reared in ways that increase substantially their risk of poor academic and social outcomes. Quite possibly, finding ways to reduce teen and single parenthood would do as much to improve the achievement of black children relative to white children as would changes in education policies or practices.

The third area for further study relates to the implications for students' achievement of various school-based strategies being promoted through the No Child Left Behind policy to raise performance of students at the bottom—for example, supplemental and compensatory services as well as accountability policies. Understanding what is working, for whom, and under what conditions to raise the performance of blacks and other low-achieving groups and knowing the profile of those children still left behind would be valuable input to future policy decisions.

The usefulness of research in all three of these areas would be enhanced if there were agreement on the essential skills children should master and how to measure achievement of them. Hopefully, this study by Murnane and others will stimulate further work on this fundamental issue.

References

- Boyd, David, and others. 2005. "Explaining the Short Careers of High-Achieving Teachers in Schools with Low-Performing Students." *American Economic Review* 95 (2): 166–71.
- Carroll, John B. 1963. "A Model of School Learning." *Teachers College Record* 64: 723–33.
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2005. "Who Teaches Whom? Race and the Distribution of Novice Teachers." *Economics of Education Review* 24: 377–92.
- . 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." Working Paper 11936. Cambridge, Mass.: National Bureau of Economic Research (January).
- Coleman, James S., and others. 1966. *Equality of Educational Opportunity*. Office of Education, National Center for Educational Statistics. Washington.
- Fryer, Ronald G., and Steven D. Levitt. 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School." *Review of Economics and Statistics* 86 (2): 447–64.
- . 2005. "The Black-White Test Score Gap through Third Grade." Working Paper 11049. Cambridge, Mass.: National Bureau of Economic Research (January).
- Hanushek, Eric A. 2003. "The Failure of Input-Based Schooling Policies." *Economic Journal* 113 (February): F64–F98.
- Hanushek, Eric A., and Steven G. Rivkin. 2006. "The Evolution of the Black-White Achievement Gap in Elementary and Middle Schools." Paper presented at the American Economic Association Annual Meetings, Boston, Massachusetts, January 6–8.
- Kane, Thomas J., and Douglas O. Staiger. 2005. "Identifying Effective Teachers with Imperfect Information." Working Paper. Harvard Graduate School of Education (April).
- Levy, Frank, and Richard J. Murnane. 2004. *The New Division of Labor: How Computers are Creating the Next Job Market*. Princeton University Press.
- Margo, Robert A. 1986. "Educational Achievement in Segregated School Systems: The Effects of 'Separate-but-Equal.'" *American Economic Review* 76 (4): 794–801.
- Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, Mass.: Ballinger.
- Murnane, Richard J., and Barbara R. Phillips. 1981. "Learning by Doing, Vintage, and Selection: Three Pieces of the Puzzle Relating Teaching Experience and Teaching Performance." *Economics of Education Review* 1 (4): 453–65.
- Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 78 (2): 251–66.
- NICHD Early Child Care Research Network. 1999. "Child Care and Mother-Child Interaction in the First Three Years of Life." *Developmental Psychology* 35: 1399–413.

- . 2001. “Nonmaternal Care and Family Factors in Early Development: An Overview of the NICHD Study of Early Child Care.” *Applied Developmental Psychology* 22 (5): 457–92.
- . 2002. “The Relation of Global First-Grade Classroom Environment to Structural Classroom Features and Teacher and Student Behaviors.” *Elementary School Journal* 102 (5): 367–87.
- . 2005. “A Day in Third Grade: A Large-Scale Study of Teacher Quality and Teacher and Student Behavior.” *Elementary School Journal* 105 (3): 305–23.
- Rasch, Georg. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica* 73 (2): 417–58.
- Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94 (2): 247–52.
- Rouse, Cecilia Elena, Jeanne Brooks-Gunn, and Sara McLanahan, eds. 2005. *The Future of Children: School Readiness: Closing Racial and Ethnic Gaps*. Brookings and Woodrow Wilson School of Public and International Affairs, Princeton University.
- Todd, Petra E., and Kenneth I. Wolpin. 2003. “On the Specification and Estimation of the Production Function for Cognitive Achievement.” *Economic Journal* 113.
- Woodcock, Richard, and M. Bunner Johnson. 1989. *Woodcock-Johnson Psycho-Educational Battery, Revised*. Allen, Texas: DLM Teaching Resources.
- Woodcock, Richard, and N. Mather. 1989. “W-J-R Tests of Achievement: Examiner’s Manual.” In *Woodcock-Johnson Psycho-Educational Battery, Revised*. Edited by Richard Woodcock and M. Bunner Johnson. Allen, Texas: DLM Teaching Resources.